Jak funguje fulltextové vyhledávání

Tomáš Hlucháň

Obsah

- Vyhledávání
- Komponenty
  - Crawler
  - Indexace
  - Výdej
  - Přepis dotazu
  - Čeština
  - Řazení
- Seznam.cz Vyhledávání

SEZNAM.CZ
...najdu tam, co hledám

Typy vyhledávání

- Sekvenční
- Hierarchie
- Index

Obsah vs. Rejstřík

Obsah vs. Rejstřík

1. Strom	1
2. Vyhledání uzlu	2
3. Vložení uzlu	3
4. Rušení uzlu	4

Rušení	4
Strom	1
Uzlu	2,3,4
Vložení	3
Vyhledání	2

Hledání dotazu

- jednoslovný dotaz
- dvouslovný dotaz
- negativní operátor

Hlavní komponenty

Crawler

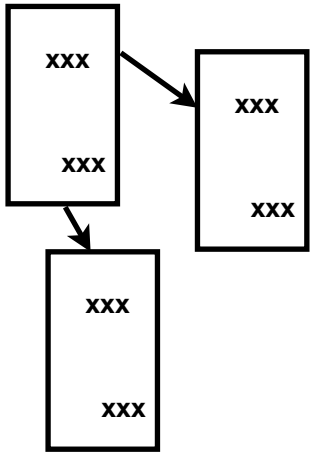
Indexace

Výdej

SEZNAM.CZ
...najdu tam, co hledám

Hlavní komponenty

Crawler

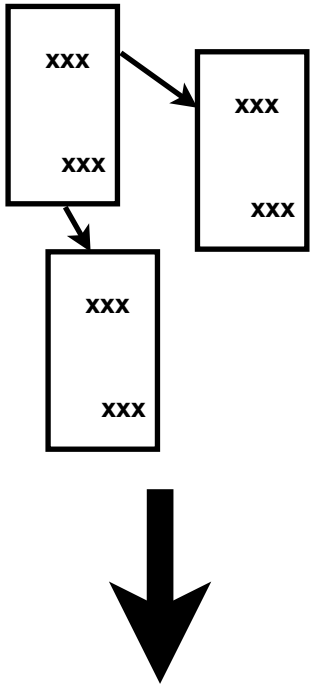


Indexace

Výdej

Hlavní komponenty

Crawler



Indexace

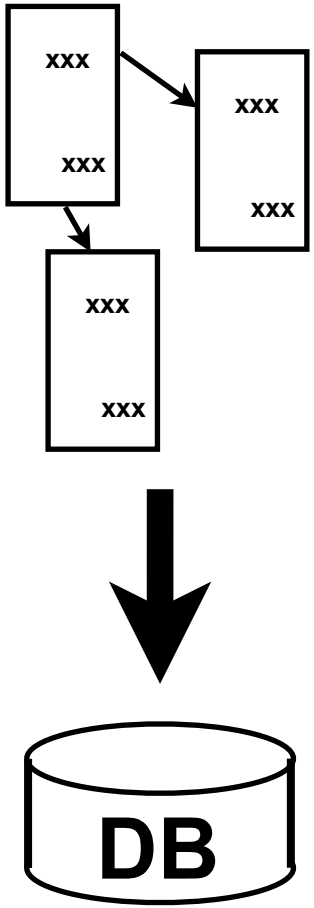
Výdej

Hlavní komponenty

Crawler

Indexace

Výdej



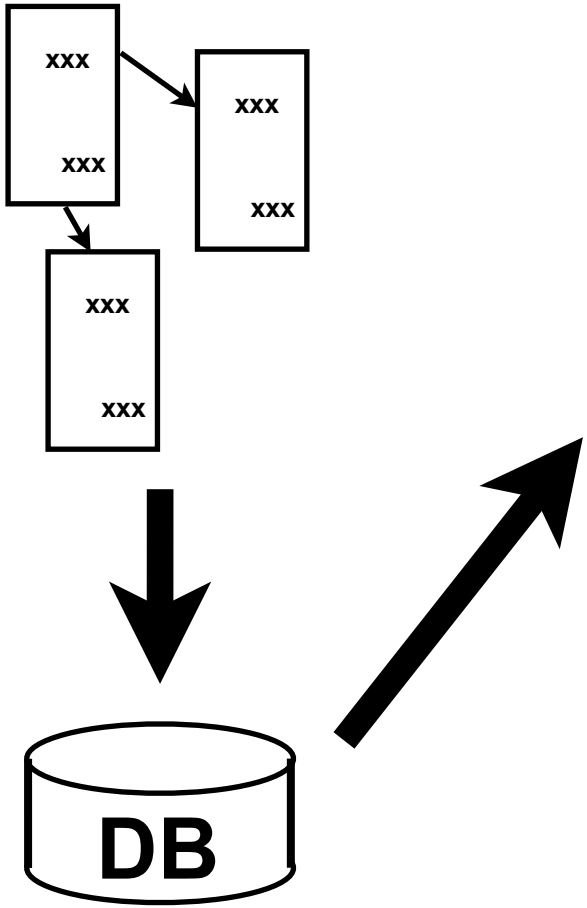
SEZNAM.CZ
...najdu tam, co hledám

Hlavní komponenty

Crawler

Indexace

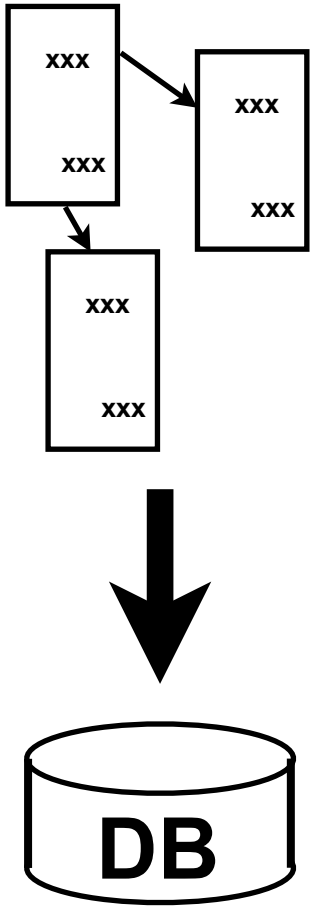
Výdej



SEZNAM.CZ
...najdu tam, co hledám

Hlavní komponenty

Crawler



Indexace

invertované indexy

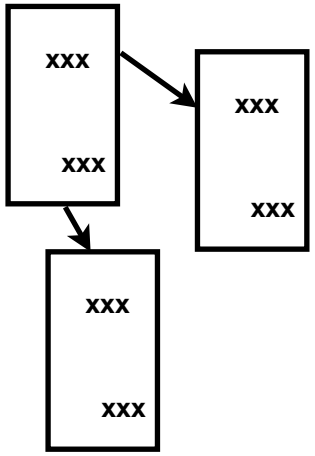
Strom
- **1,1,...**

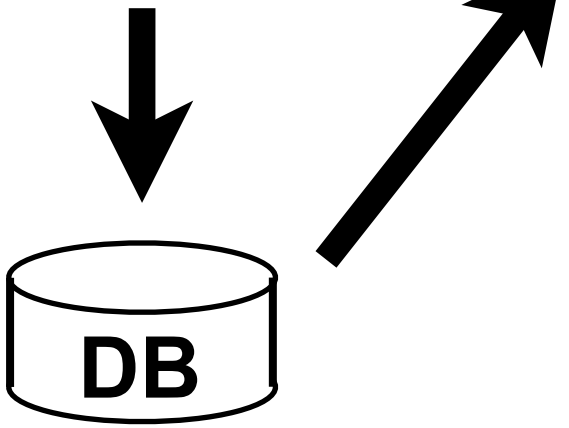
Uzlu
- **2,2,...**
- **3,2,...**
- **4,2,....**

Výdej

Hlavní komponenty

Crawler





Indexace

invertované indexy

Strom
- **1,1,...**

Uzlu
- **2,2,...**
- **3,2,...**
- **4,2,....**

dokumenty

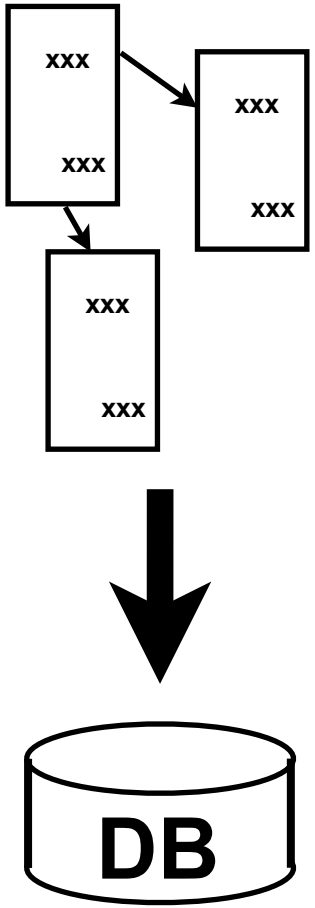
- **1, rank, jazyk**
- **2, rank, jazyk**
- **3...**

Výdej

SEZNAM.CZ
...najdu tam, co hledám

Hlavní komponenty

Crawler



Indexace

invertované indexy

Strom
- **1,1,...**

Uzlu
- **2,2,...**
- **3,2,...**
- **4,2,....**

dokumenty

- **1, rank, jazyk**
- **2, rank, jazyk**
- **3...**

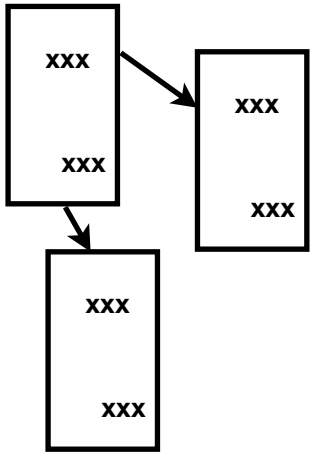
Výdej

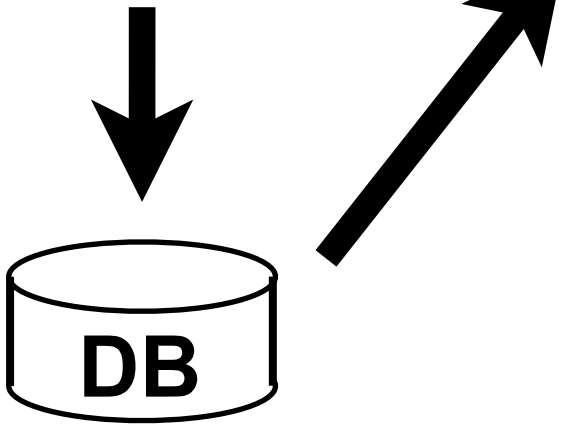
- analýza dotazu

SEZNAM.CZ
...najdu tam, co hledám

Hlavní komponenty

Crawler





Indexace

invertované indexy

Strom
- **1,1,...**

Uzlu
- **2,2,...**
- **3,2,...**
- **4,2,....**

dokumenty

- **1, rank, jazyk**
- **2, rank, jazyk**
- **3...**

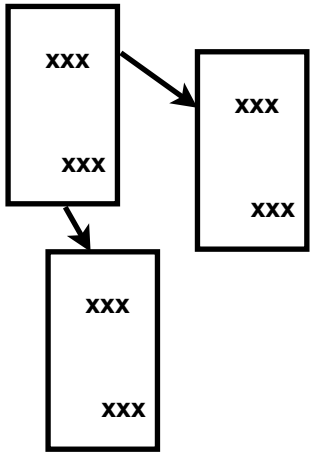
Výdej

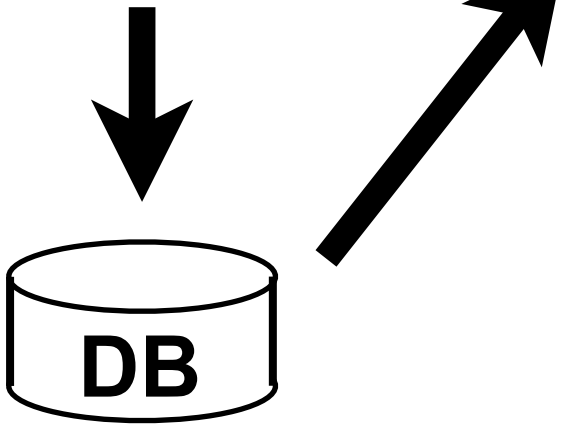
- analýza dotazu
- vyhledání

SEZNAM.CZ
...najdu tam, co hledám

Hlavní komponenty

Crawler





Indexace

invertované indexy

Strom
- **1,1,...**

Uzlu
- **2,2,...**
- **3,2,...**
- **4,2,....**

dokumenty

- **1, rank, jazyk**
- **2, rank, jazyk**
- **3...**

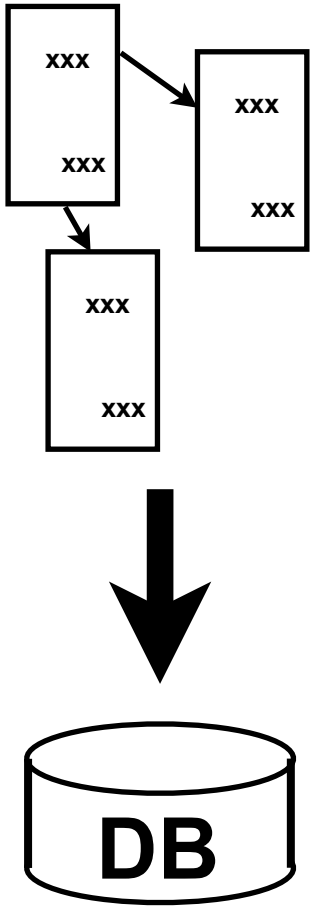
Výdej

- analýza dotazu
- vyhledání
- řazení dle relevance

SEZNAM.CZ
...najdu tam, co hledám

Hlavní komponenty

Crawler



Indexace

invertované indexy

Strom
- **1,1,...**

Uzlu
- **2,2,...**
- **3,2,...**
- **4,2,....**

dokumenty

- **1, rank, jazyk**
- **2, rank, jazyk**
- **3...**

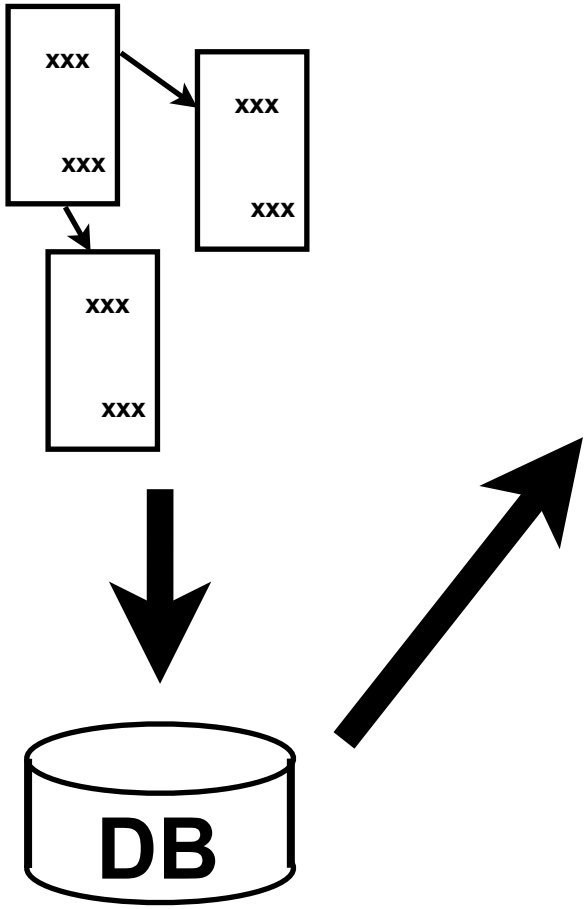
Výdej

- analýza dotazu
- vyhledání
- řazení dle relevance
- snipety

SEZNAM.CZ
...najdu tam, co hledám

Hlavní komponenty

Crawler



Indexace

invertované indexy

Strom
- **1,1,...**

Uzlu
- **2,2,...**
- **3,2,...**
- **4,2,....**

dokumenty

- **1, rank, jazyk**
- **2, rank, jazyk**
- **3...**

texty

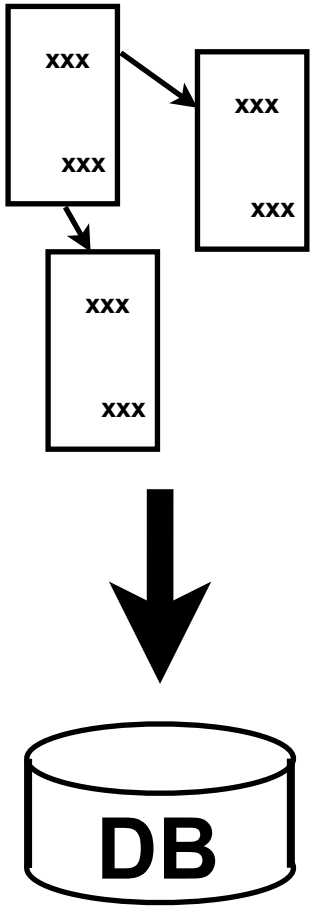
- **1, text stránky**
- **2, text ..**

Výdej

- analýza dotazu
- vyhledání
- řazení dle relevance
- snipety

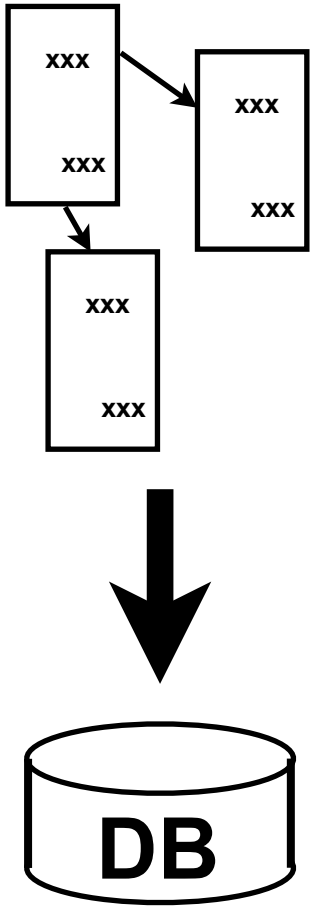
SEZNAM.CZ
...najdu tam, co hledám

Crawler



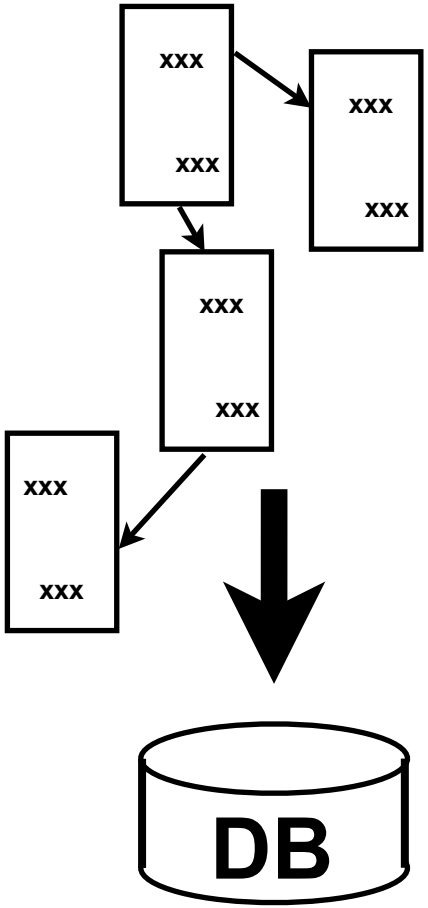
SEZNAM.CZ
...najdu tam, co hledám

Crawler



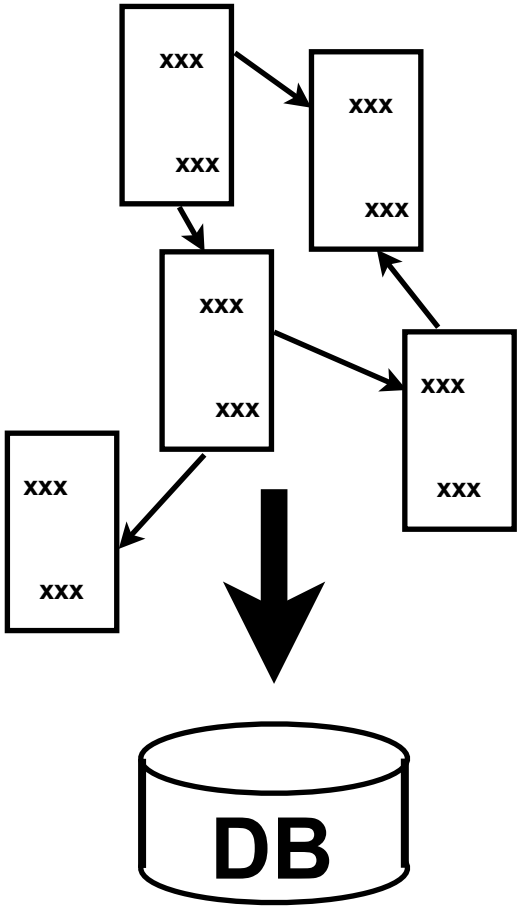
- Co vybrat?

Crawler



- Co vybrat?

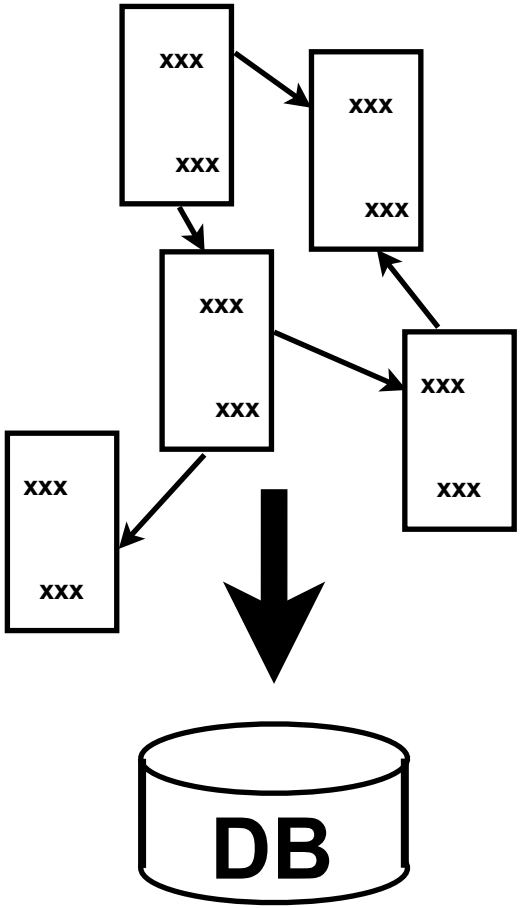
Crawler



- Co vybrat?

SEZNAM.CZ
...najdu tam, co hledám

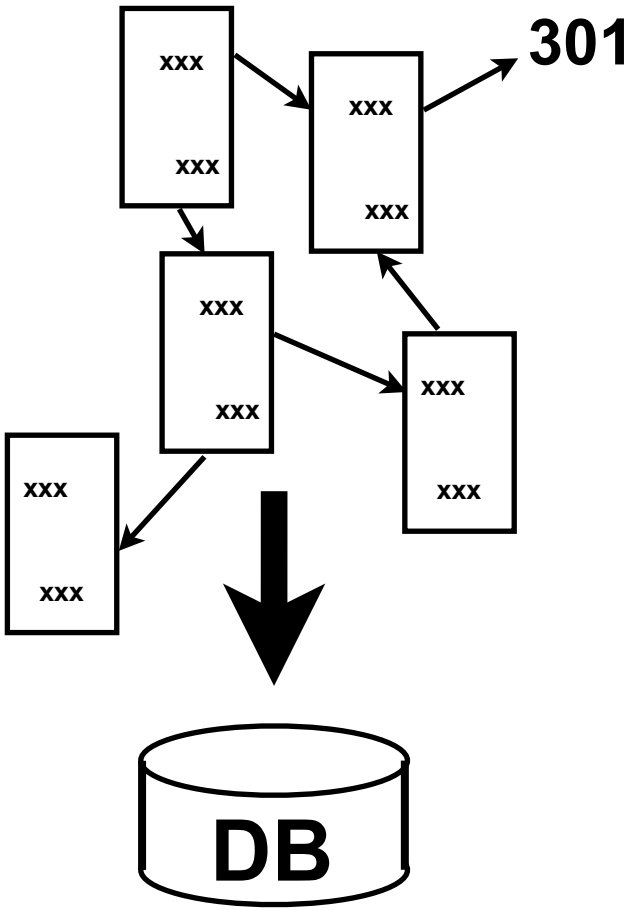
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript

SEZNAM.CZ
...najdu tam, co hledám

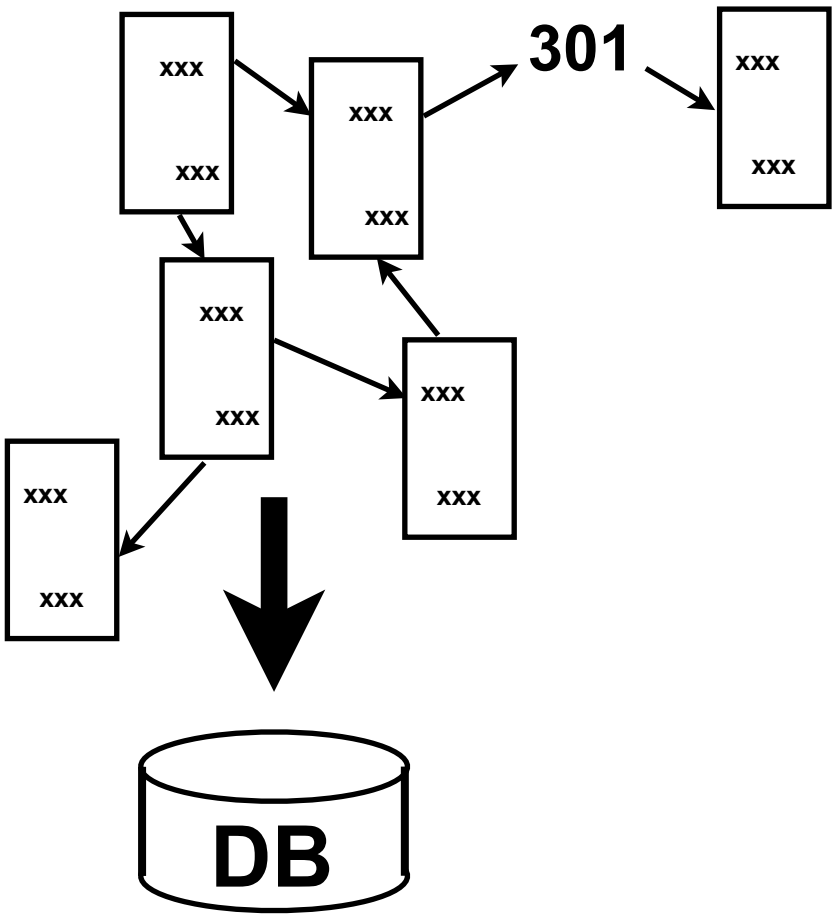
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript

SEZNAM.CZ
...najdu tam, co hledám

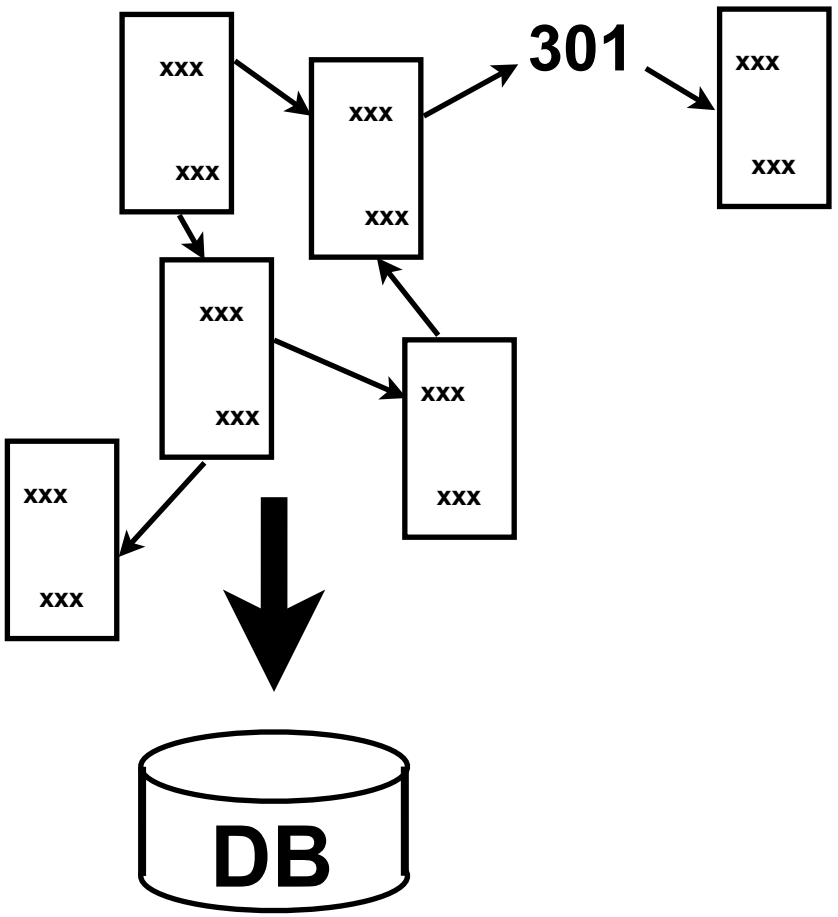
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript

SEZNAM.CZ
...najdu tam, co hledám

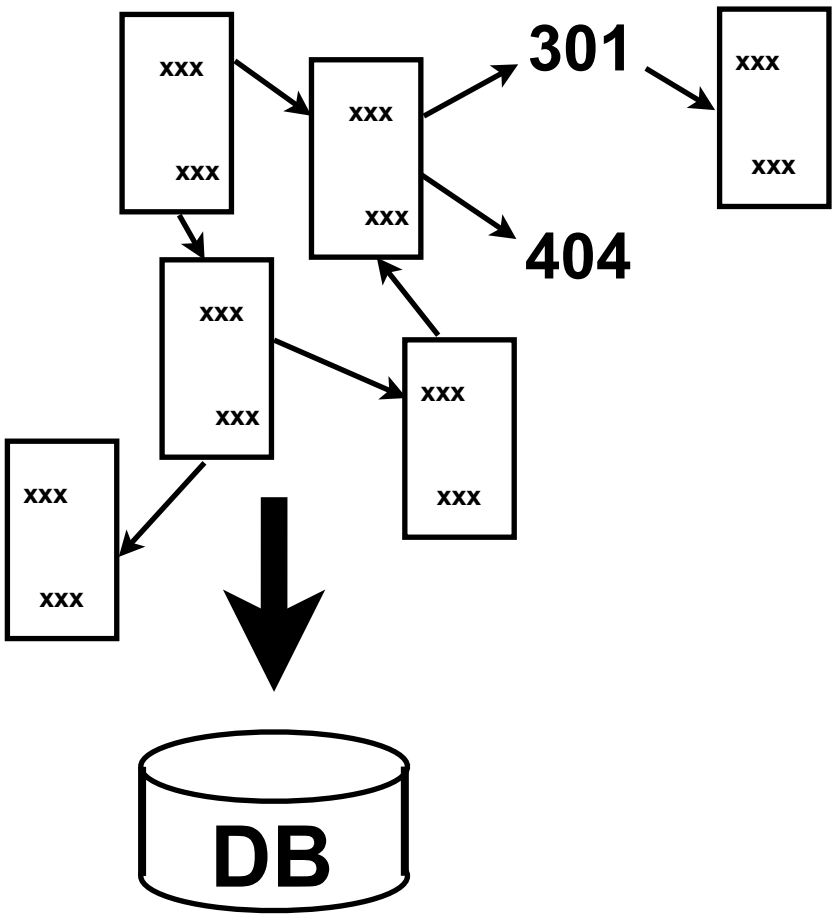
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript
- Chyby
  - 404
  - 500
  - zahlcení sítě

SEZNAM.CZ
...najdu tam, co hledám

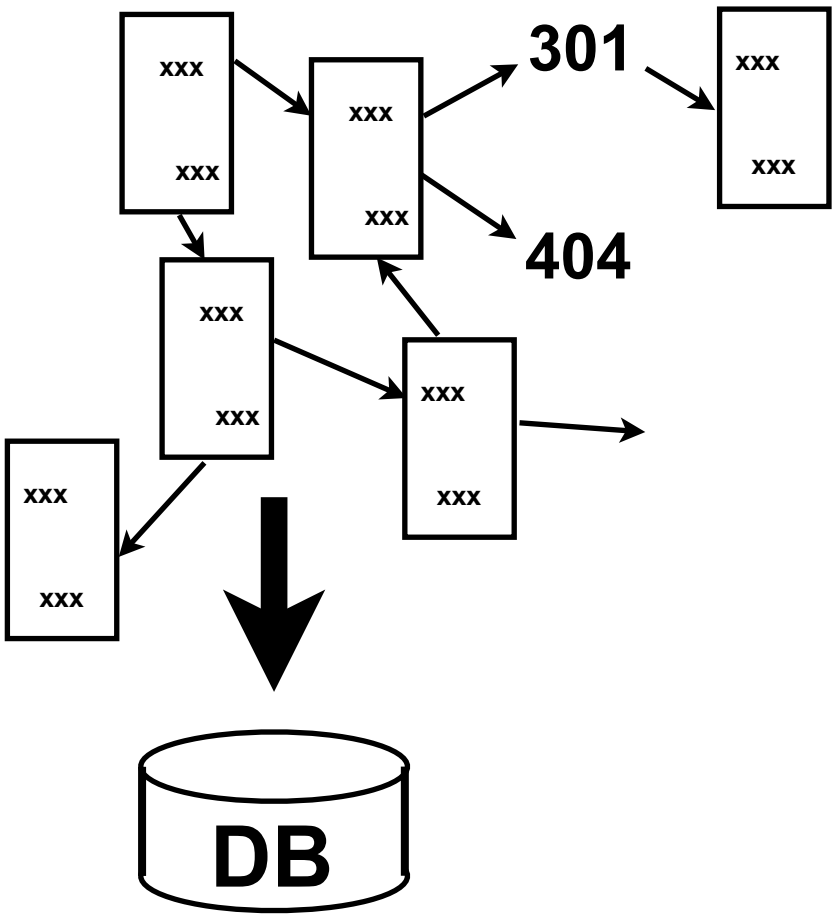
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript
- Chyby
  - 404
  - 500
  - zahlcení sítě

SEZNAM.CZ
...najdu tam, co hledám

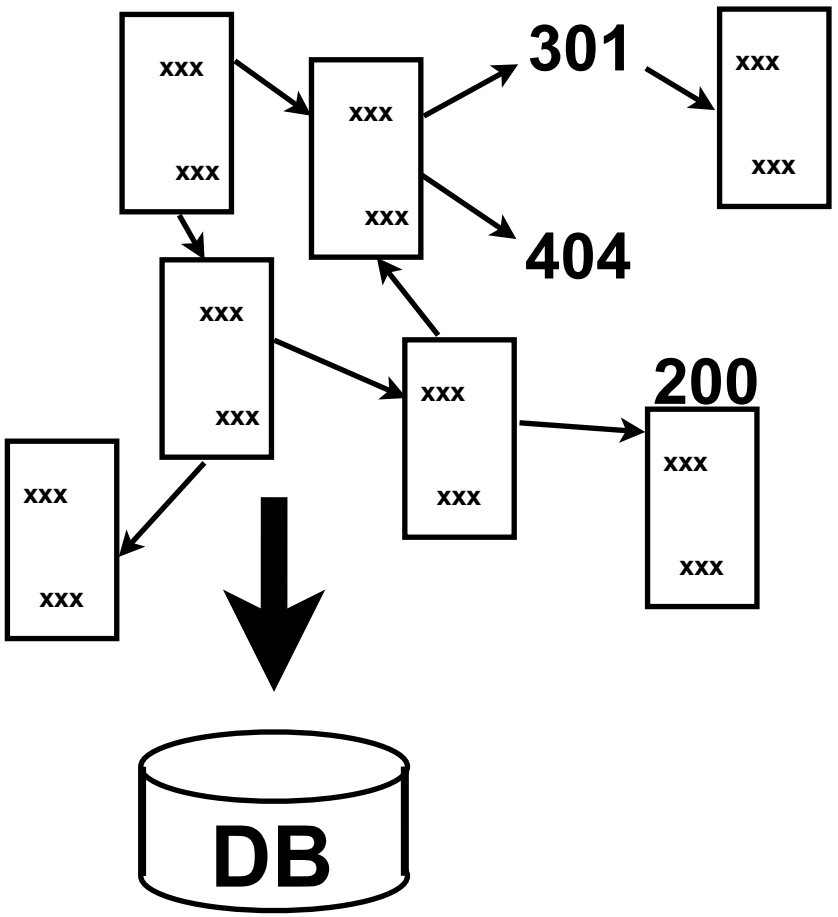
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript
- Chyby
  - 404
  - 500
  - zahlcení sítě

SEZNAM.CZ
...najdu tam, co hledám

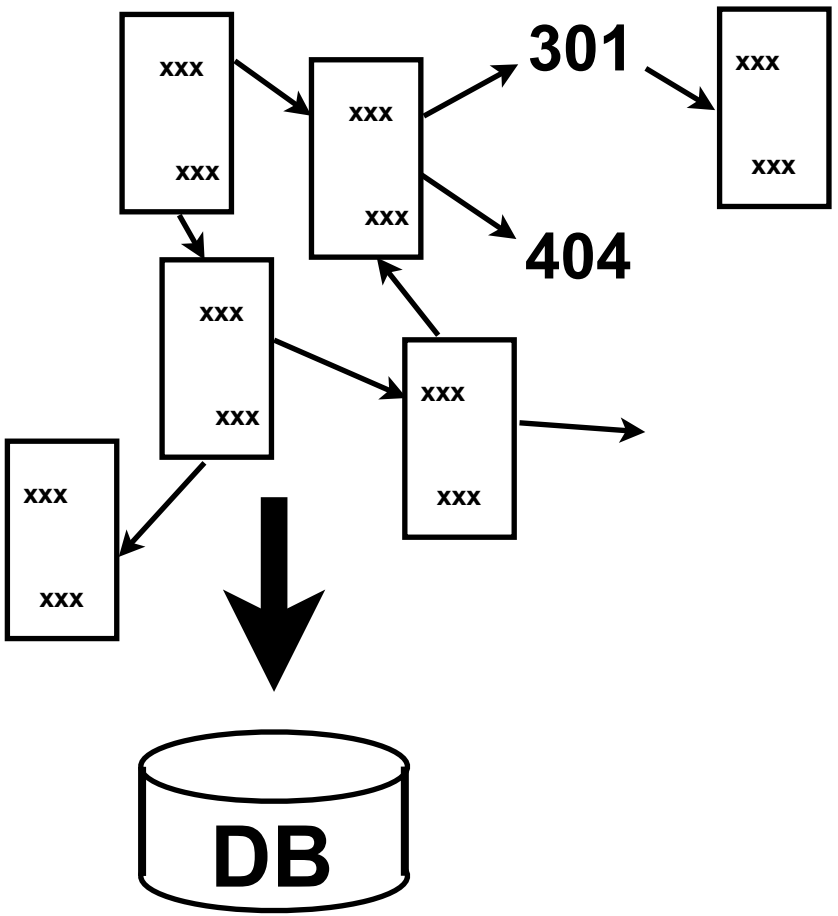
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript
- Chyby
  - 404
  - 500
  - zahlcení sítě

SEZNAM.CZ
...najdu tam, co hledám

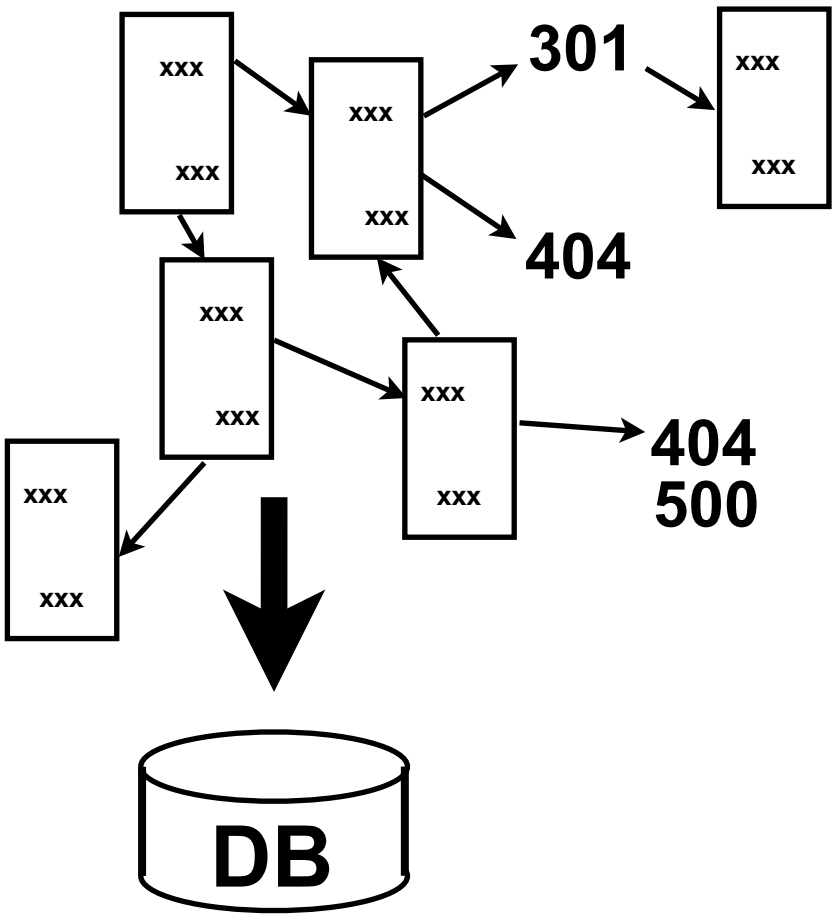
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript
- Chyby
  - 404
  - 500
  - zahlcení sítě

SEZNAM.CZ
...najdu tam, co hledám

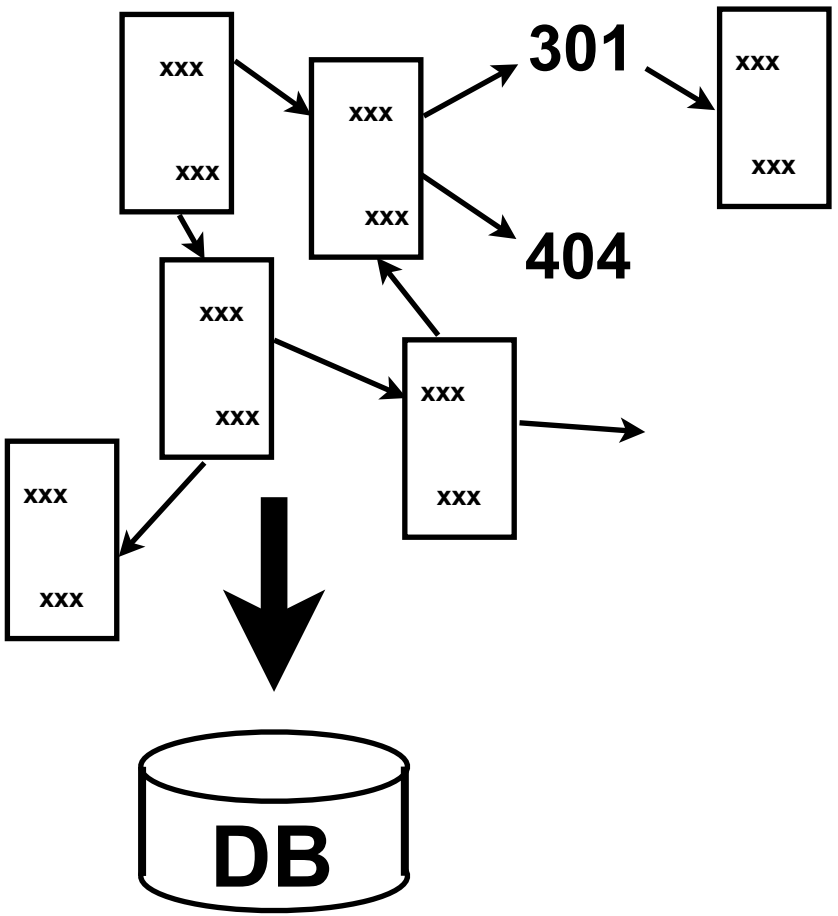
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript
- Chyby
  - 404
  - 500
  - zahlcení sítě

SEZNAM.CZ
...najdu tam, co hledám

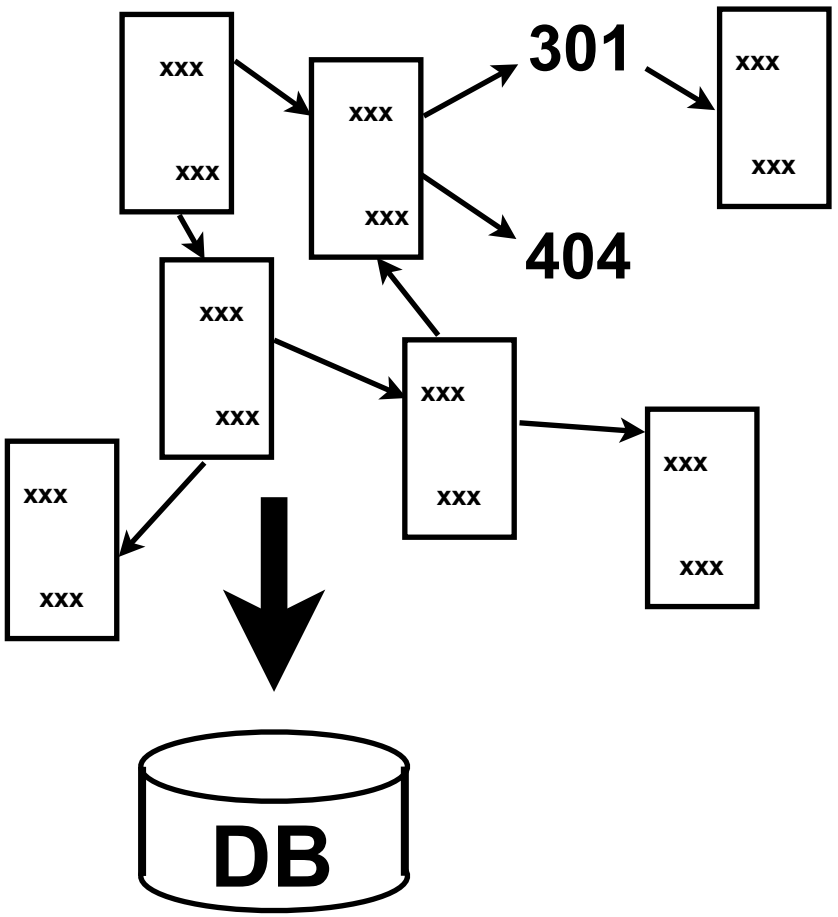
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript
- Chyby
  - 404
  - 500
  - zahlcení sítě

SEZNAM.CZ
...najdu tam, co hledám

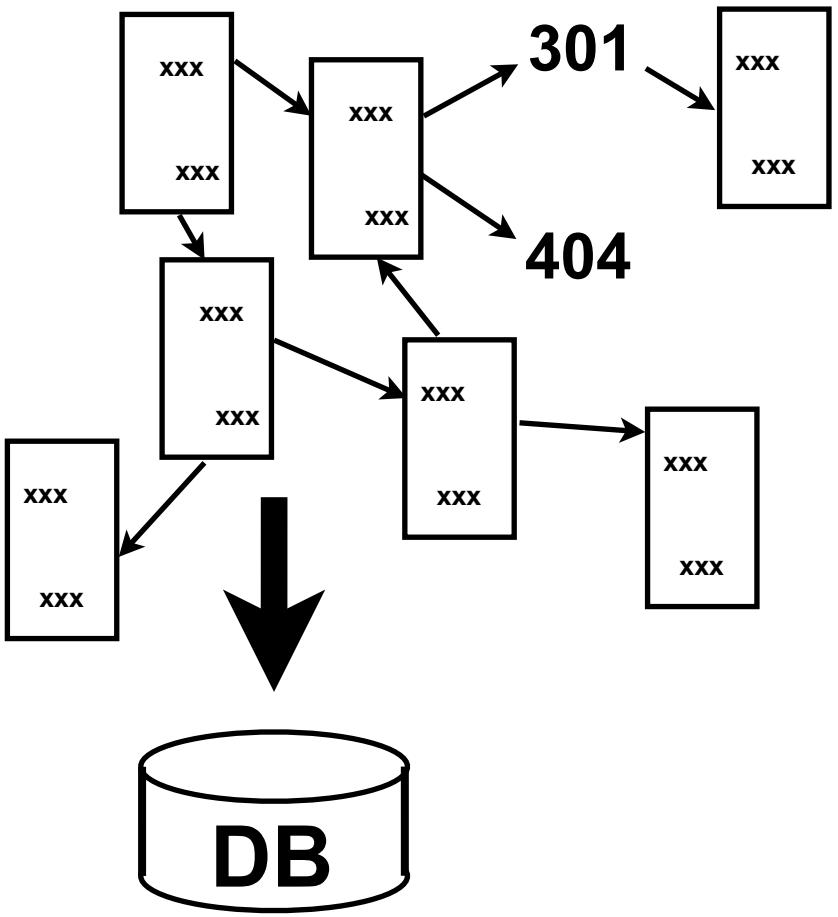
Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript
- Chyby
  - 404
  - 500
  - zahlcení sítě

SEZNAM.CZ
...najdu tam, co hledám

Crawler



- Co vybrat?
- Přesměrování
  - 301
  - javascript
- Chyby
  - 404
  - 500
  - zahlcení sítě
- Duplicita

SEZNAM.CZ
...najdu tam, co hledám

Crawler - duplicita

- Co je to duplicita?
- Úplná a částečná duplicita
- Přirozená duplicita
  - http://www.seznam.cz/ http://www.seznam.cz/index.html
  - test.php?x=1&y=2 test.php?y=2&x=1
  - Kanonizace
- Nepřirozené duplicity

Crawler - aktualizace

- Frekvence změny
  - detekce
  - predikce změny
- Důležitost
  - ranky
- RSS
- Technické limity

Crawler - DB

SEZNAM.CZ
...najdu tam, co hledám

Crawler - DB

- Obsah
  - “texty”
  - metadata
    - jazyk
    - ranky
    - odkazy
    - ....

Crawler - DB

- Obsah
  - “texty”
  - metadata
    - jazyk
    - ranky
    - odkazy
    - ....
- NoSQL

SEZNAM.CZ
...najdu tam, co hledám

Crawler - DB

- Obsah
  - "texty"
  - metadata
    - jazyk
    - ranky
    - odkazy
    - ....
- NoSQL
- Map-Reduce

SEZNAM.CZ
...najdu tam, co hledám

Crawler - DB

- Obsah
  - "texty"
  - metadata
    - jazyk
    - ranky
    - odkazy
    - ....
- NoSQL
- Map-Reduce

SEZNAM.CZ
...najdu tam, co hledám

Crawler - dodatky

SEZNAM.CZ
...najdu tam, co hledám

Crawler - dodatky

- robots.txt

SEZNAM.CZ
...najdu tam, co hledám

Crawler - dodatky

- robots.txt
- sitemapy

Crawler - dodatky

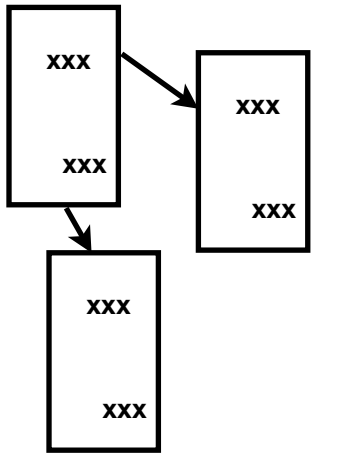
- robots.txt
- sitemapy
- ignorované parametry

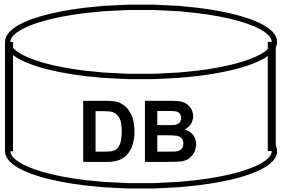
Crawler - dodatky

- robots.txt
- sitemapy
- ignorované parametry
- výběrové funkce
  - založit
  - smazat
  - indexovat

Crawler - dodatky

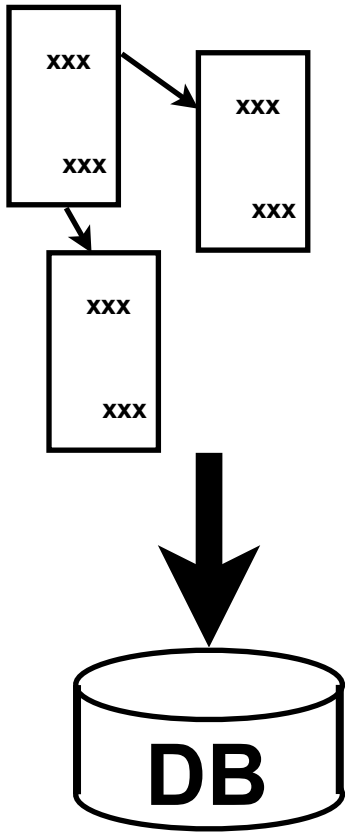
- robots.txt
- sitemapy
- ignorované parametry
- výběrové funkce
  - založit
  - smazat
  - indexovat





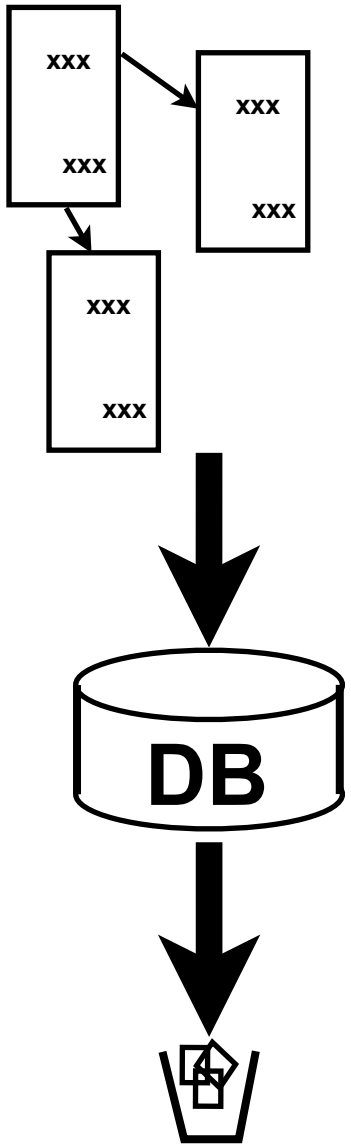
Crawler - dodatky

- robots.txt
- sitemapy
- ignorované parametry
- výběrové funkce
  - založit
  - smazat
  - indexovat



Crawler - dodatky

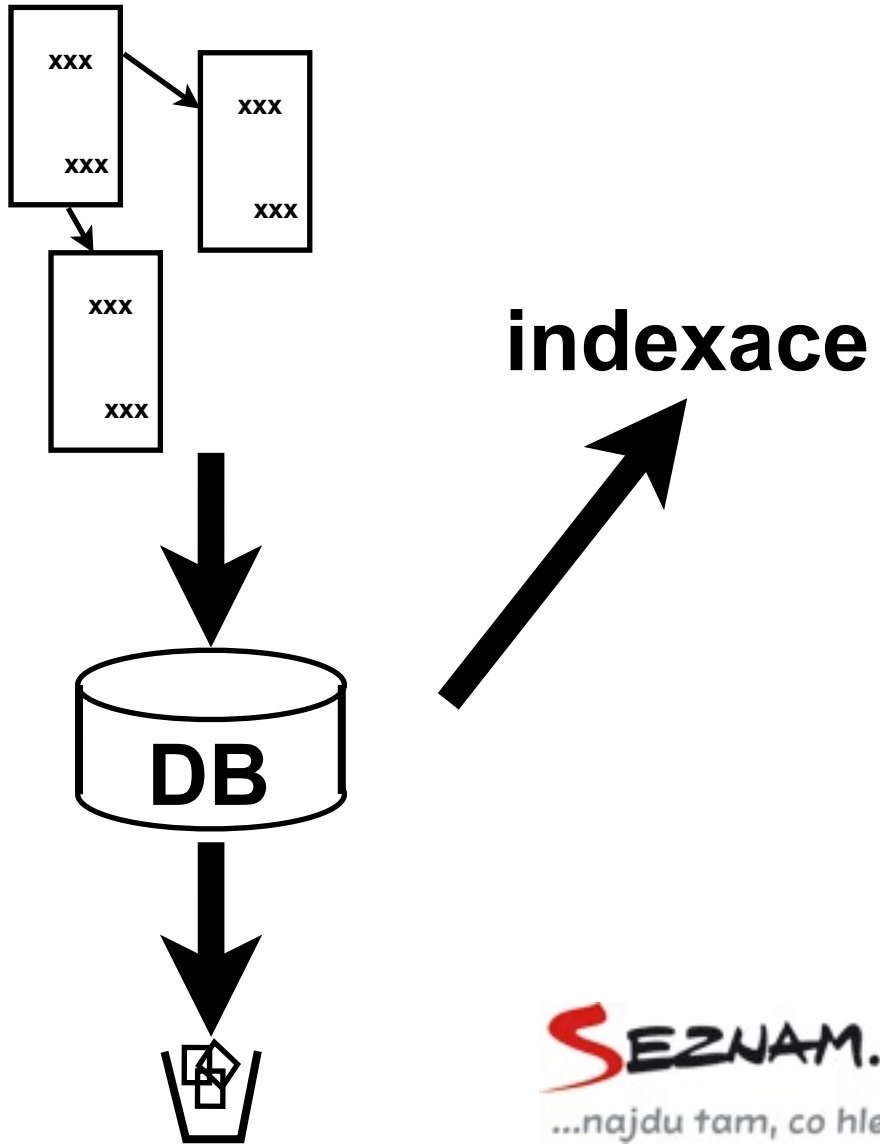
- robots.txt
- sitemapy
- ignorované parametry
- výběrové funkce
  - založit
  - smazat
  - indexovat



SEZNAM.CZ
...najdu tam, co hledám

Crawler - dodatky

- robots.txt
- sitemapy
- ignorované parametry
- výběrové funkce
  - založit
  - smazat
  - indexovat



SEZNAM.CZ
...najdu tam, co hledám

Indexace

- Získání stránky
- Převedení na “prostý” text
- Extrakce odkazů a metadat
- Výpočty metadat
- Budování rejstříku

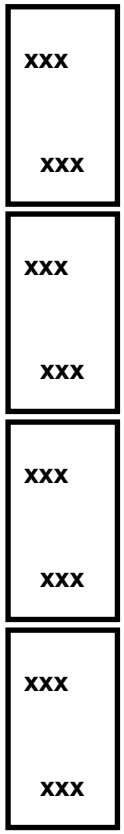
- Aktualizace indexu
  - přírůstkové
  - najednou
- Uložení odkazů a metadat

SEZNAM.CZ
...najdu tam, co hledám

Indexace - merge & split

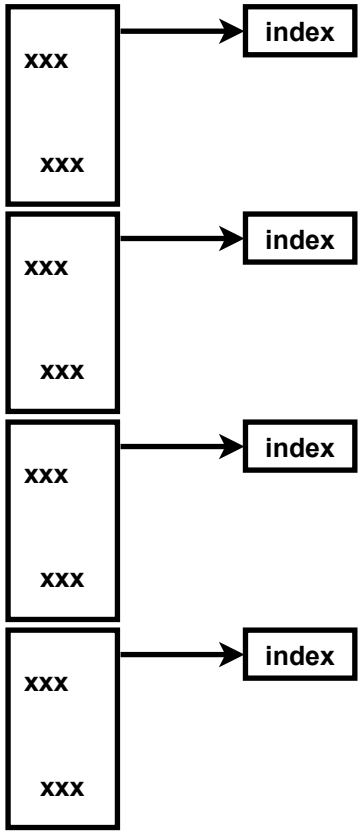
SEZNAM.CZ
...najdu tam, co hledám

Indexace - merge & split



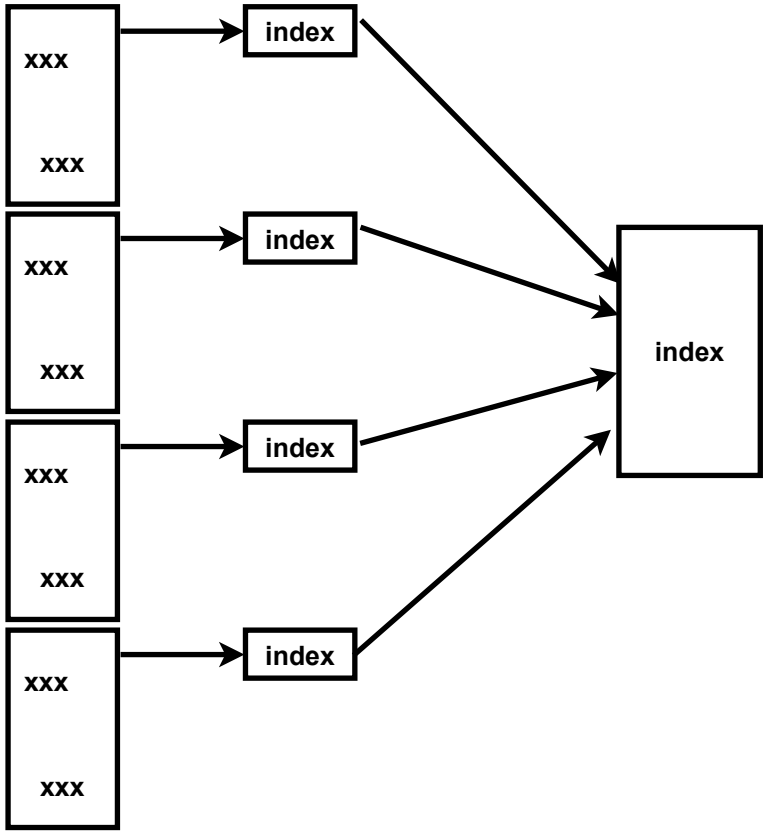
SEZNAM.CZ
...najdu tam, co hledám

Indexace - merge & split



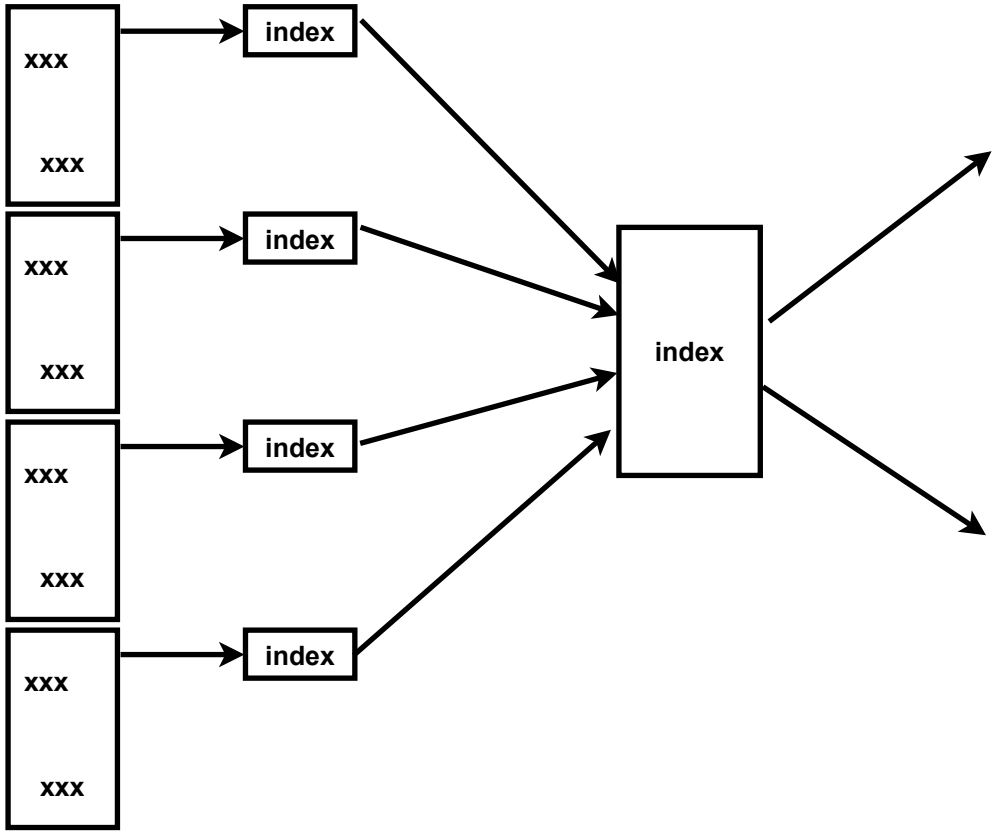
SEZNAM.CZ
...najdu tam, co hledám

Indexace - merge & split



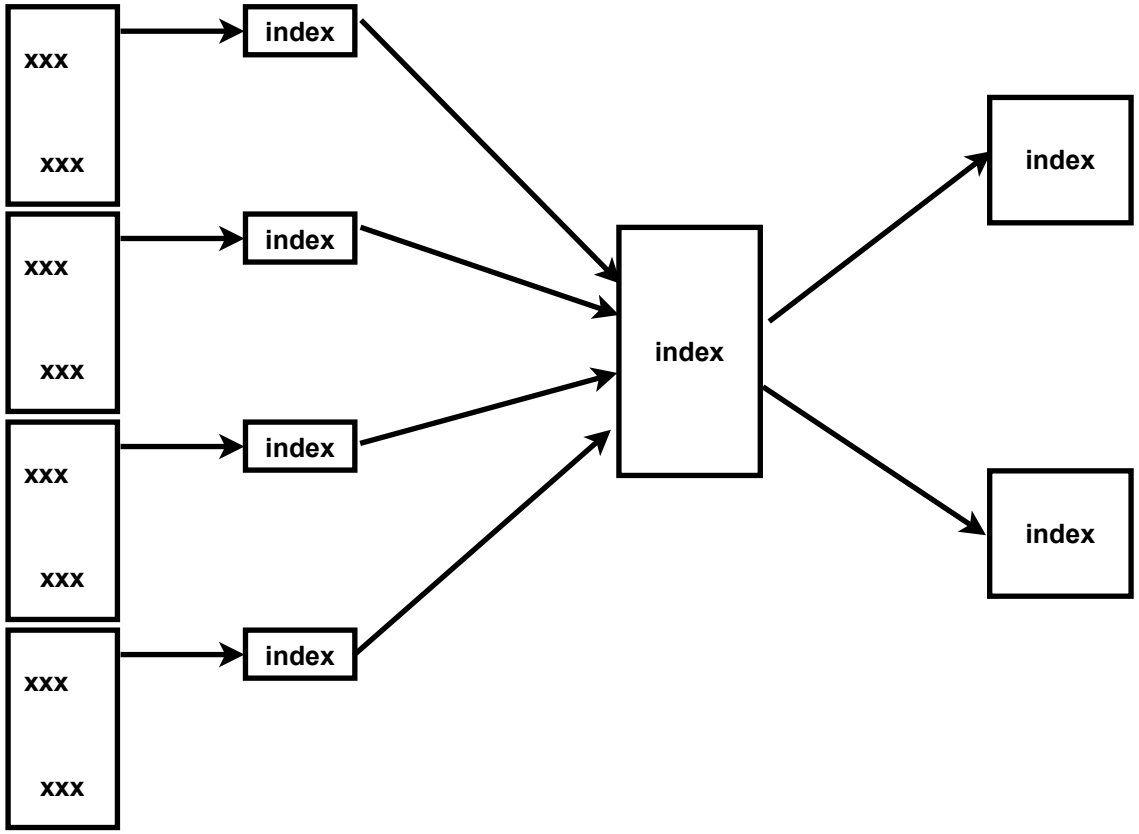
SEZNAM.CZ
...najdu tam, co hledám

Indexace - merge & split



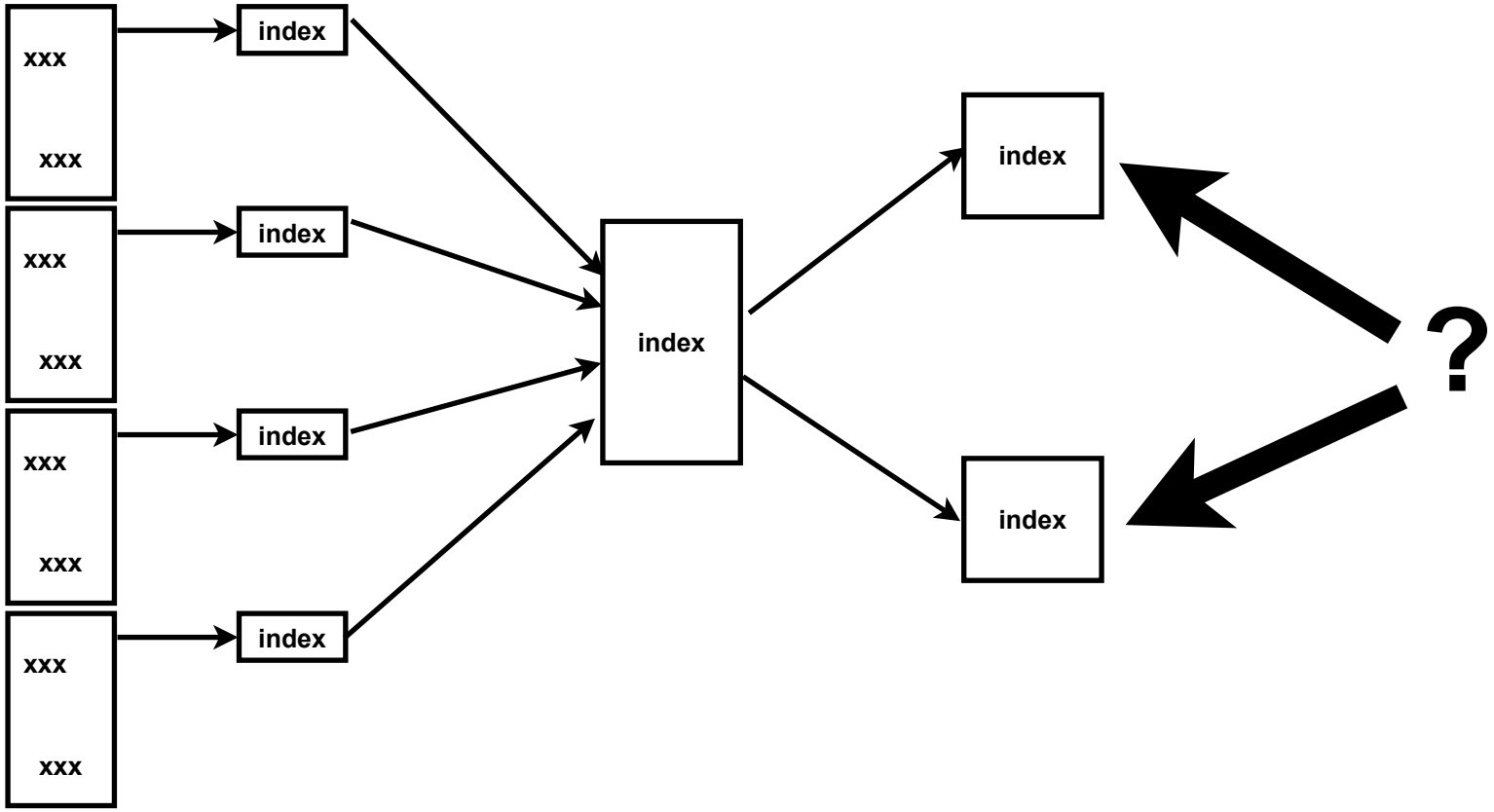
SEZNAM.CZ
...najdu tam, co hledám

Indexace - merge & split

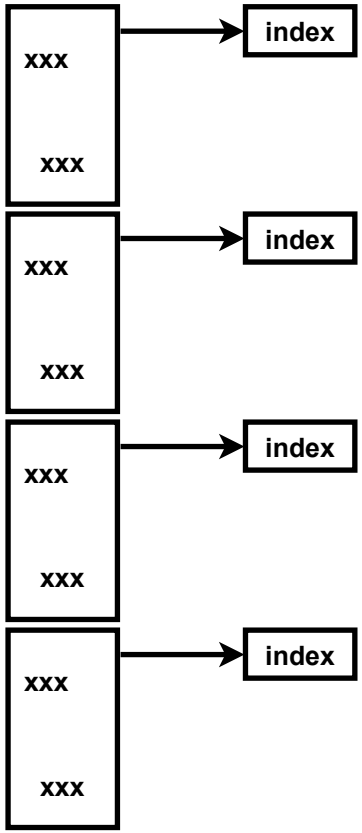


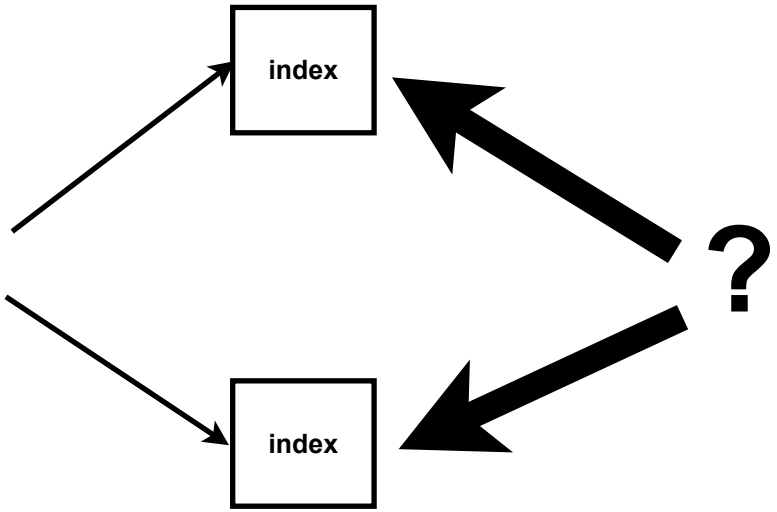
SEZNAM.CZ
...najdu tam, co hledám

Indexace - merge & split



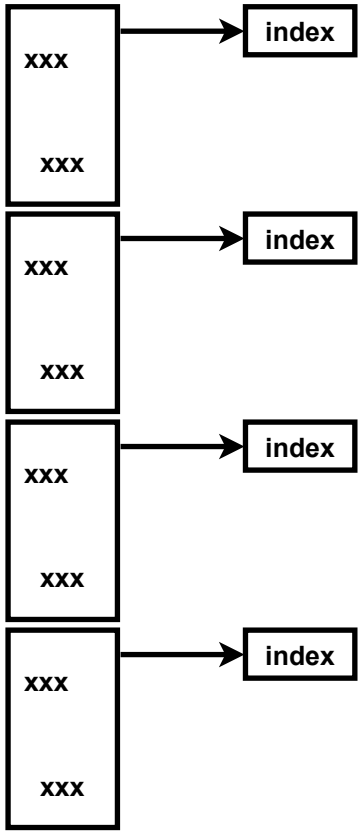
Indexace - merge & split

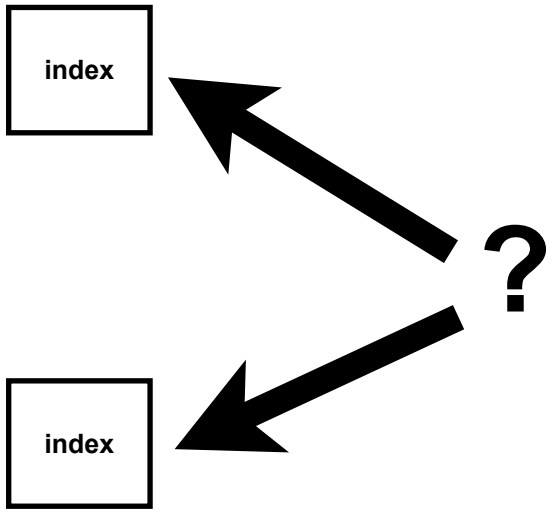




SEZNAM.CZ
...najdu tam, co hledám

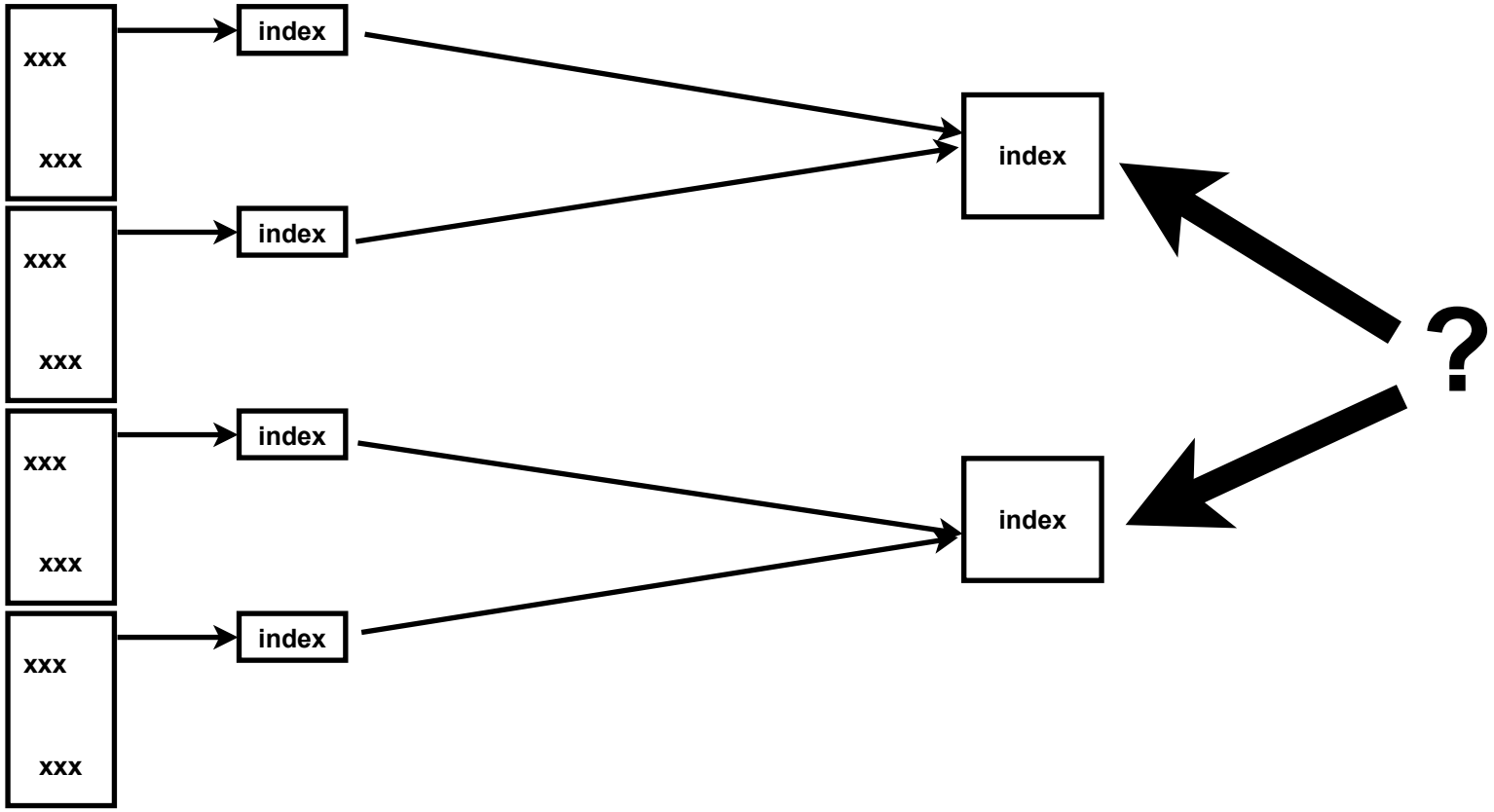
Indexace - merge & split



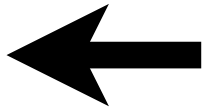


SEZNAM.CZ
...najdu tam, co hledám

Indexace - merge & split



Výdej

index  dotaz

Výdej

index ← **dotaz**

auto

Doc	Pos
1	34,23
3	12
21	45,67
34	1
67	4

Výdej

index ← **dotaz**

auto

Doc	Pos
1	34,23
3	12
21	45,67
34	1
67	4

[auto]

Výdej

index ← **dotaz**

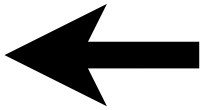
auto

Doc	Pos
1	34,23
3	12
21	45,67
34	1
67	4

[auto]

[bílé auto]

Výdej



index **dotaz**

auto

Doc	Pos
1	34,23
3	12
21	45,67
34	1
67	4

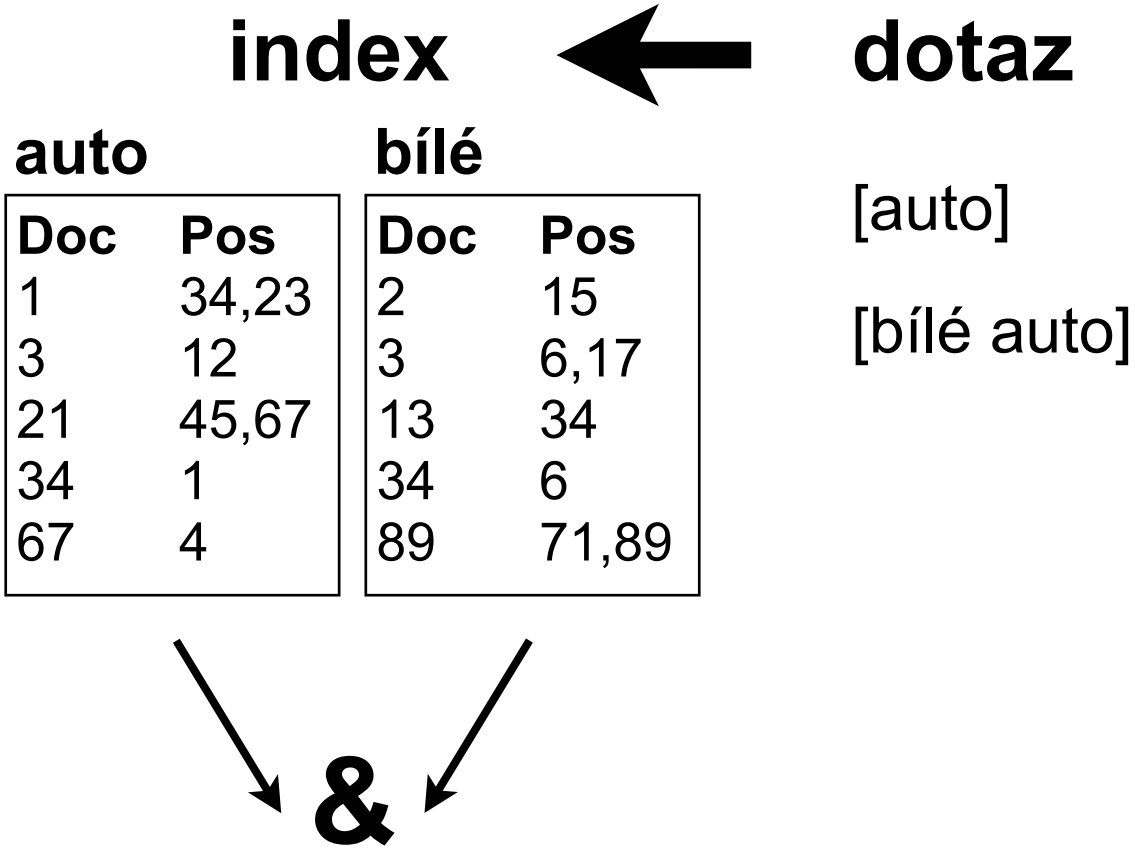
bílé

Doc	Pos
2	15
3	6,17
13	34
34	6
89	71,89

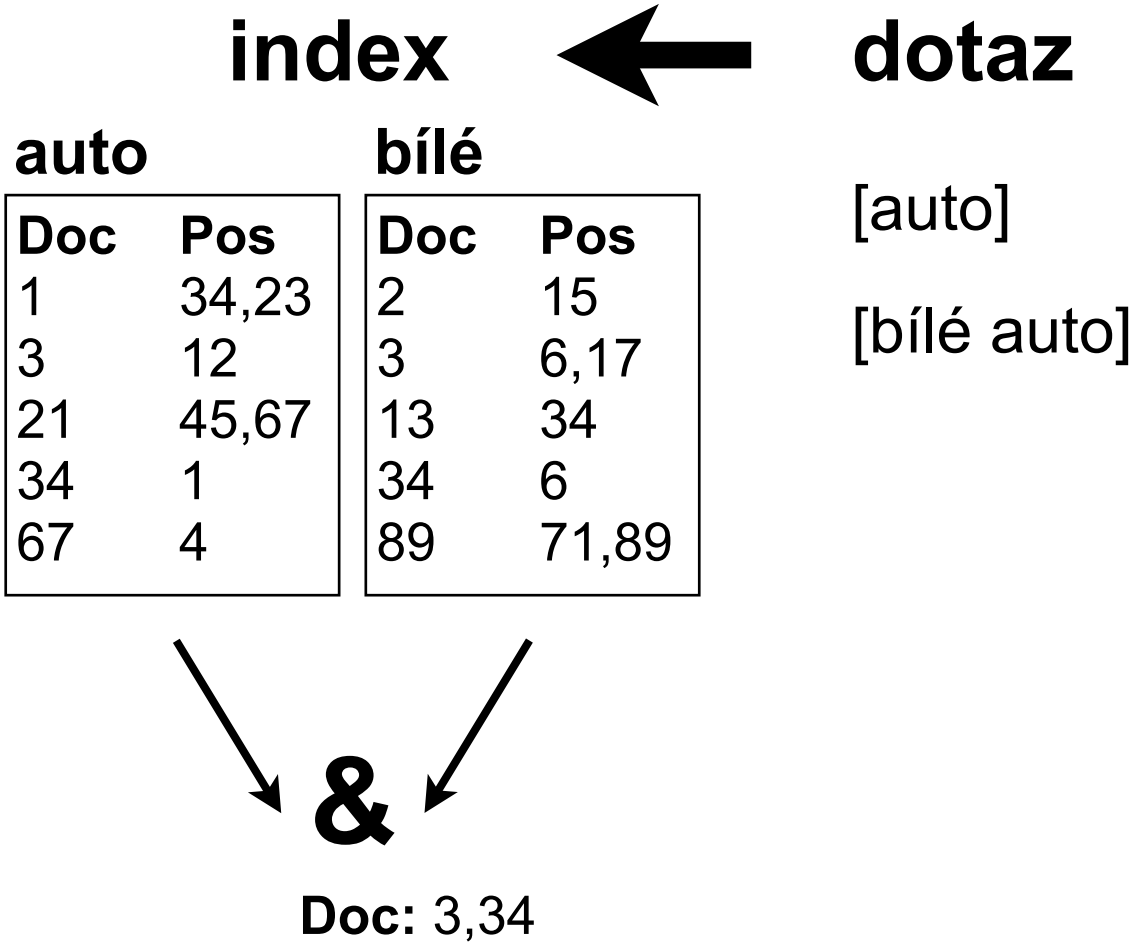
[auto]

[bílé auto]

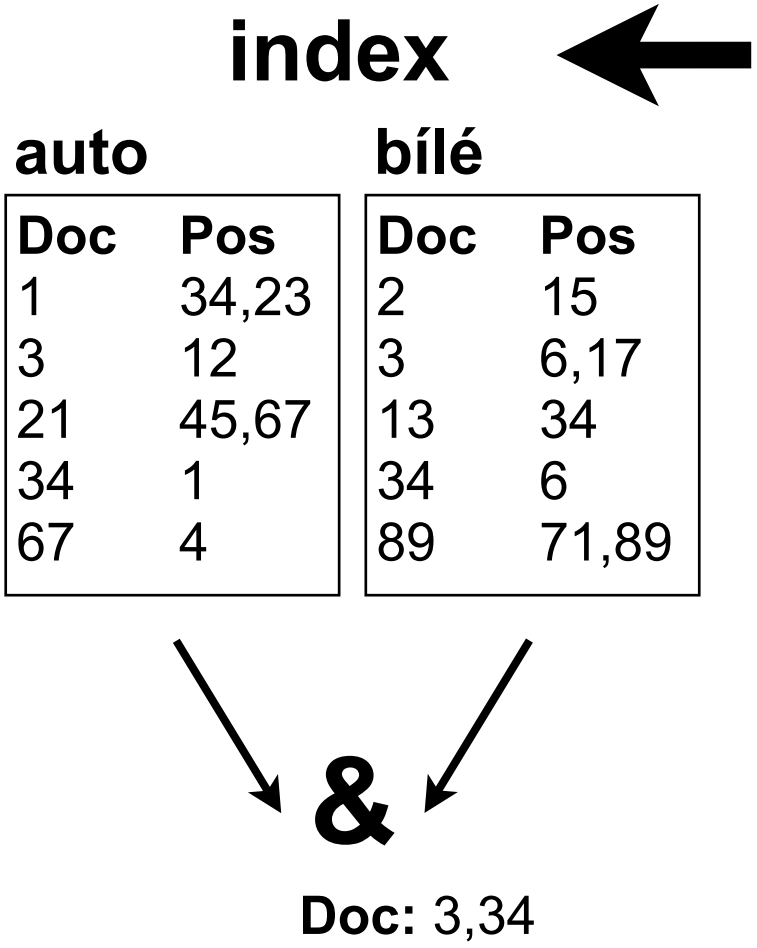
Výdej



Výdej



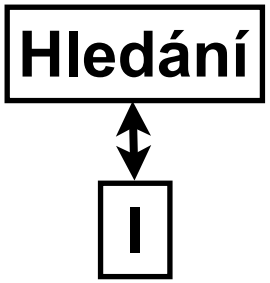
Výdej



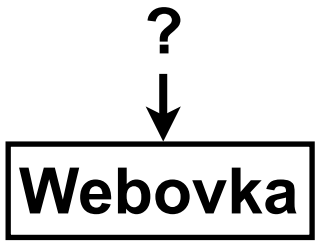
- Pochopit
- **Najít**
- Relevance
- Seřadit

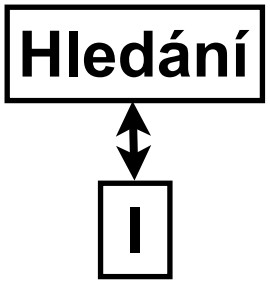
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



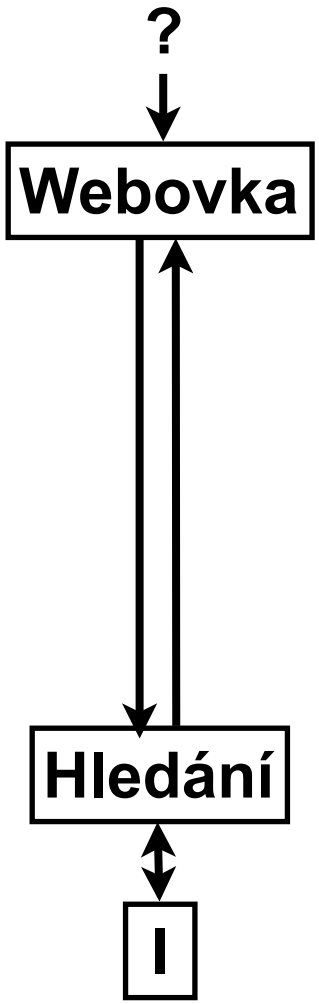
Výdej - architektura





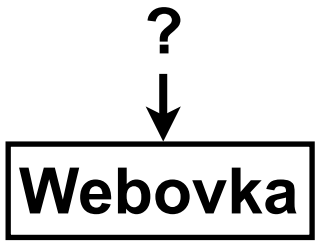
SEZNAM.CZ
...najdu tam, co hledám

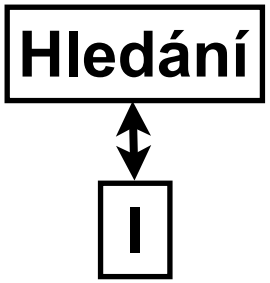
Výdej - architektura



SEZNAM.CZ
...najdu tam, co hledám

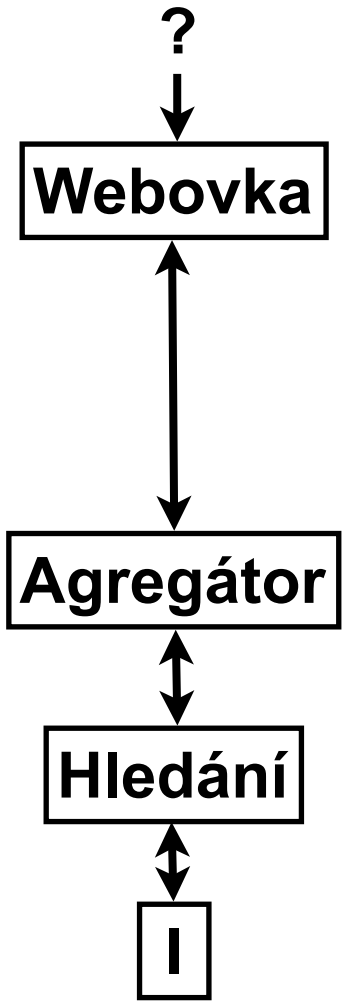
Výdej - architektura





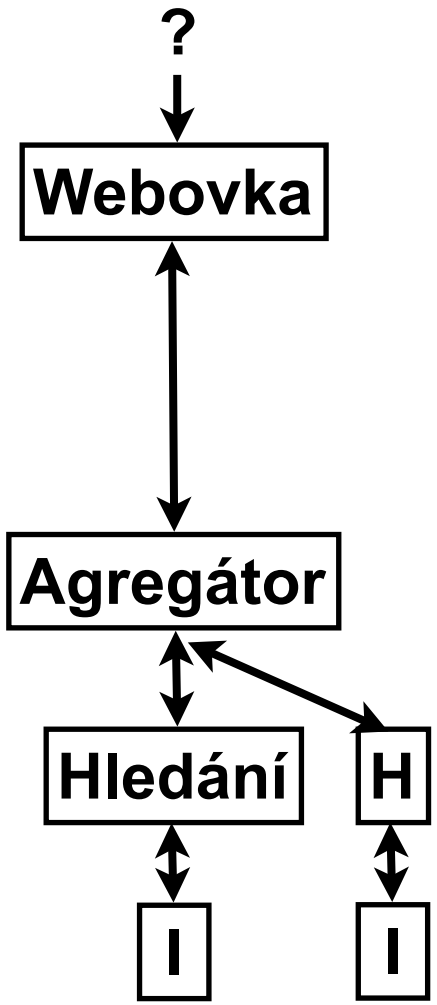
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



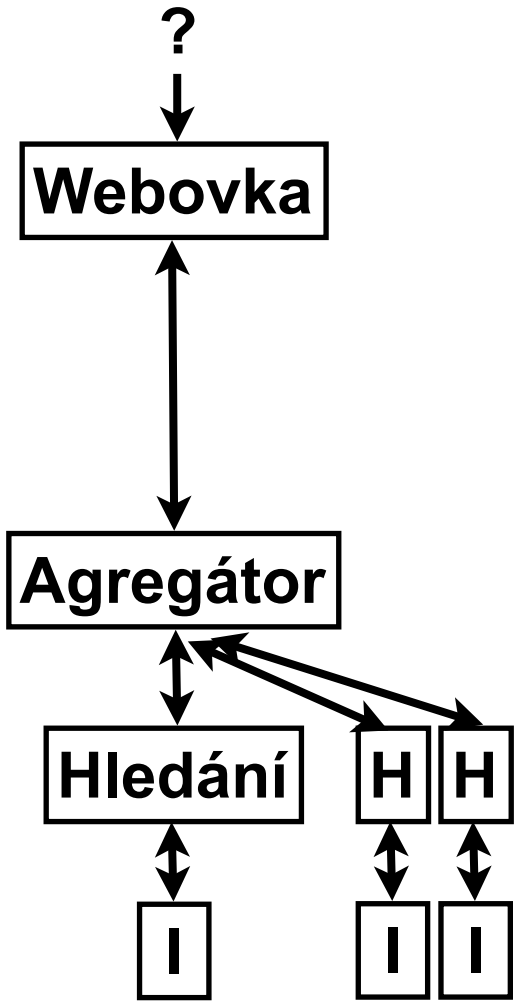
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



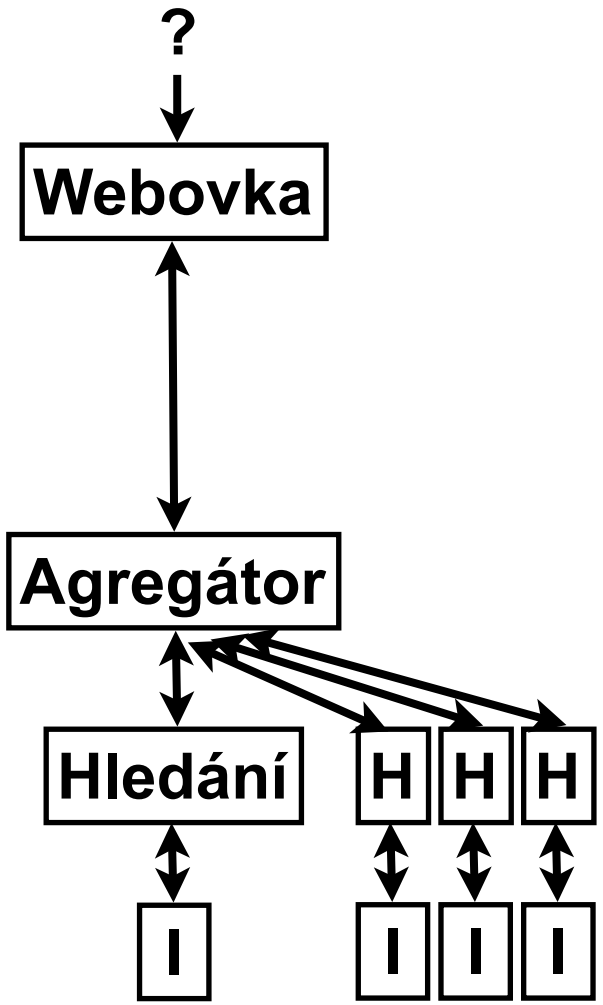
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



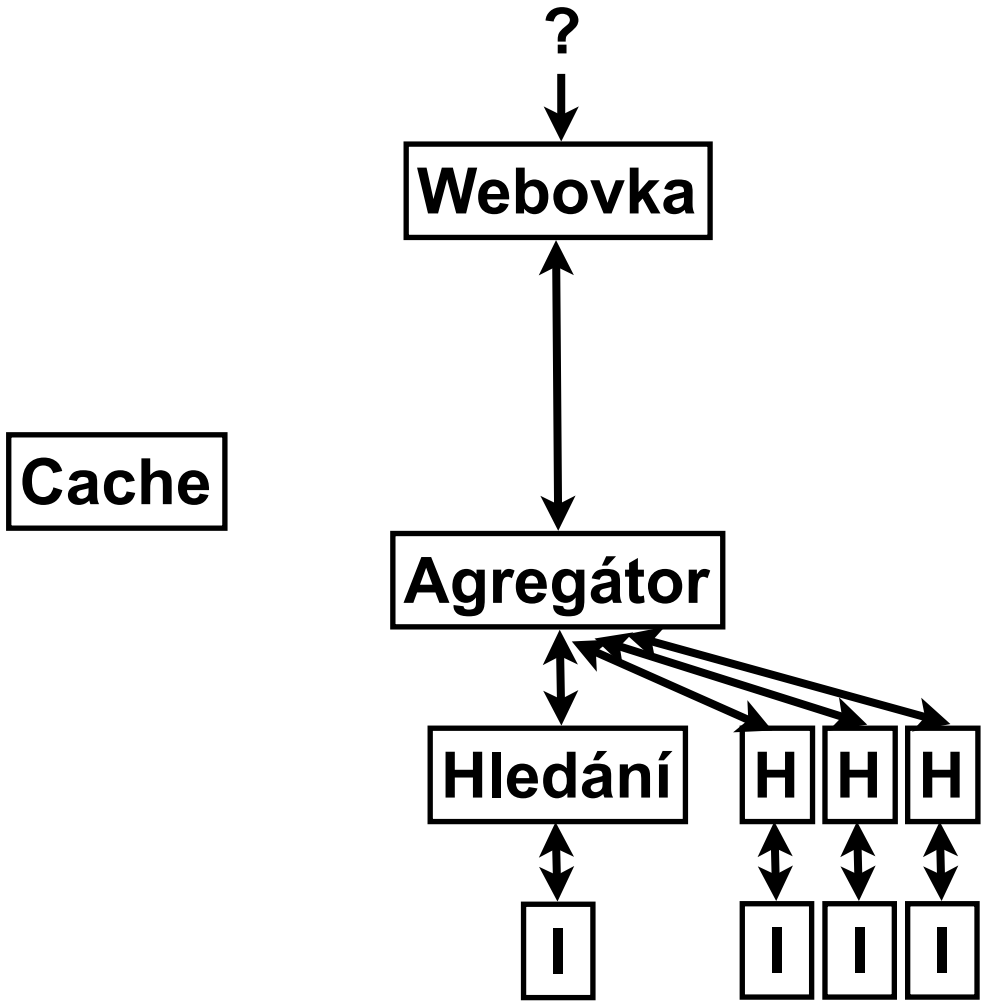
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



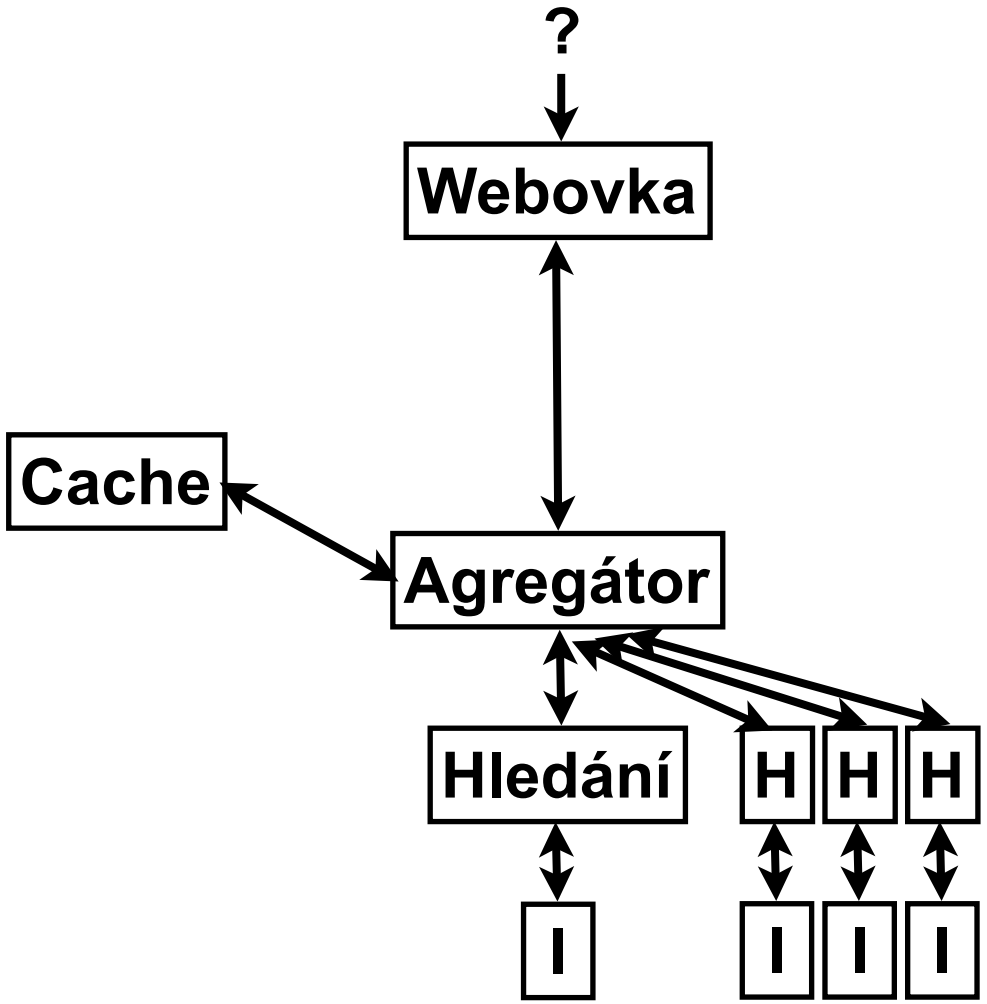
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura

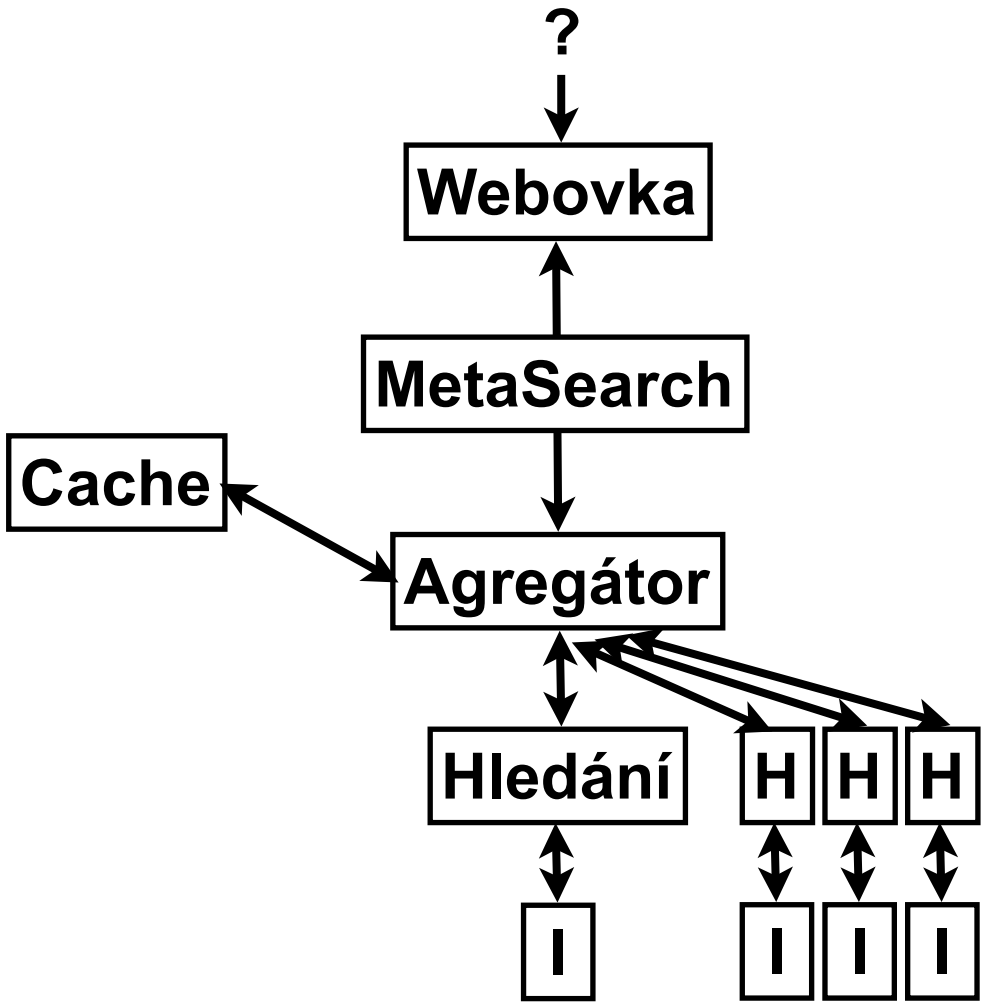


SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura

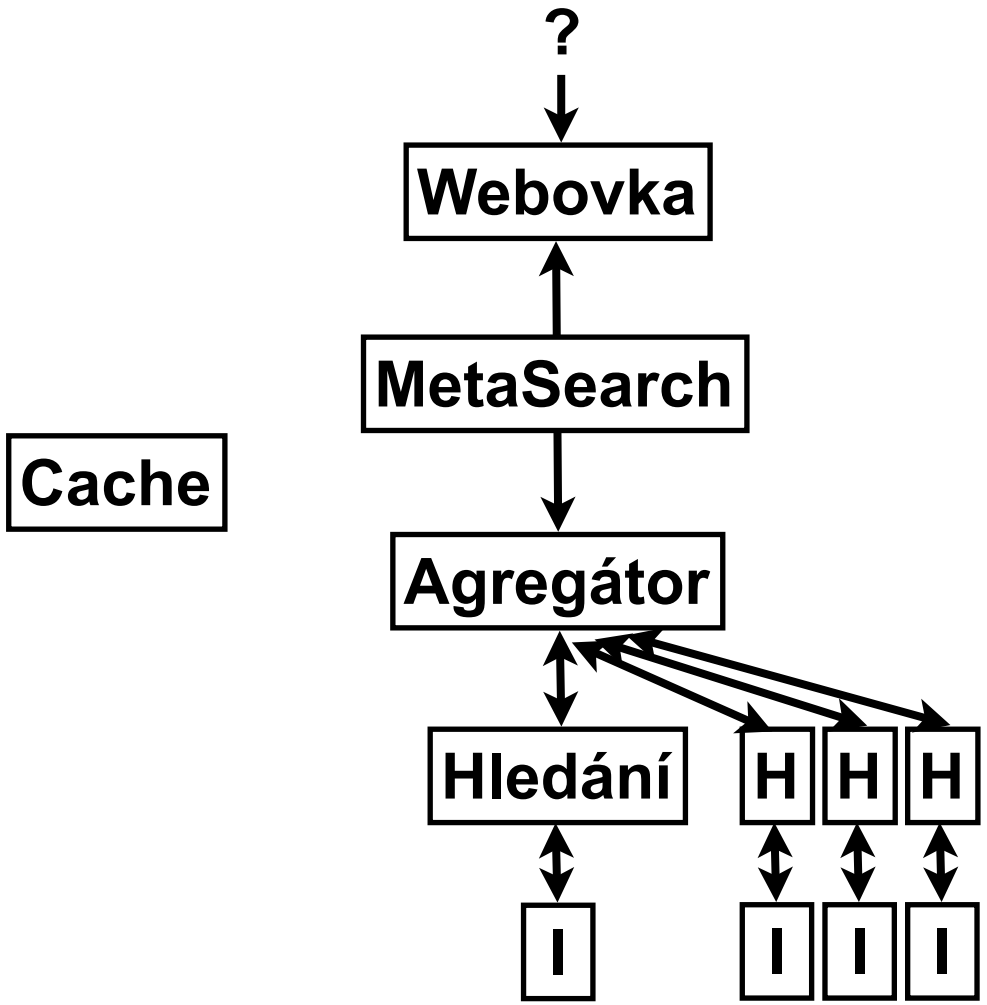


Výdej - architektura



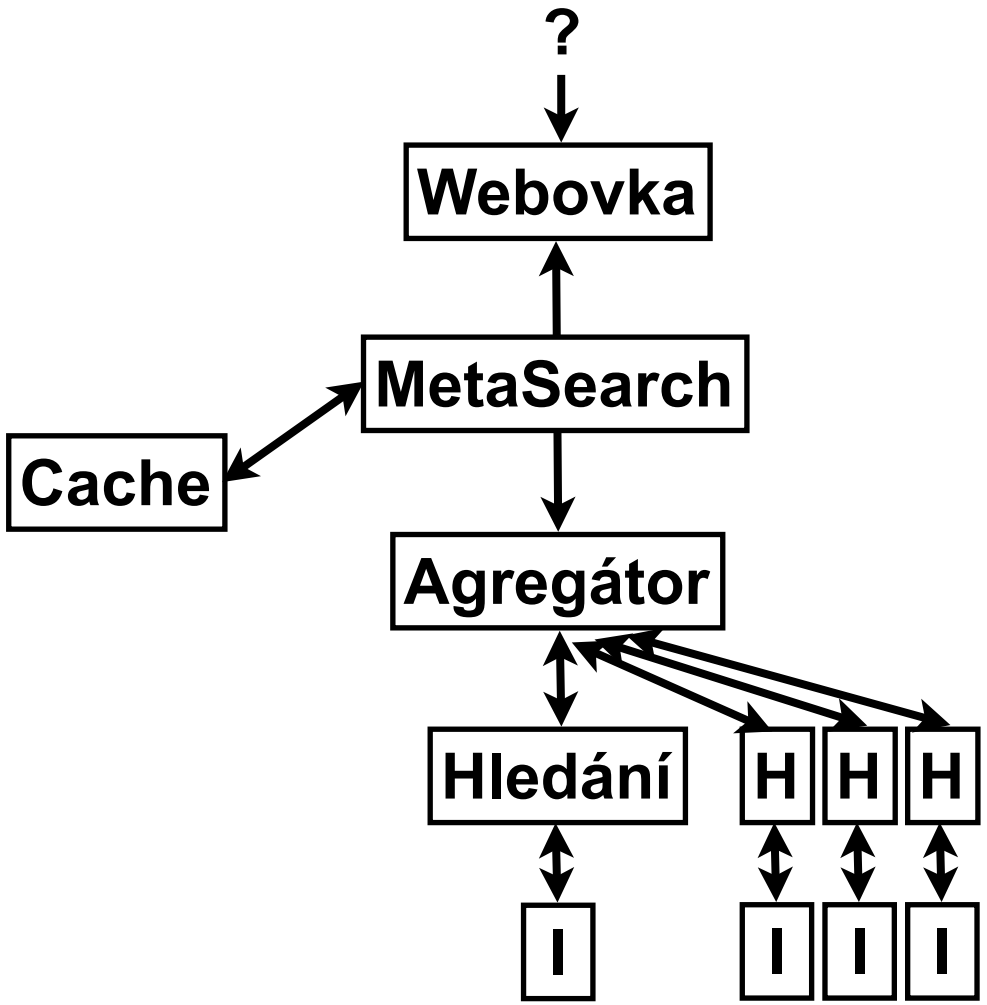
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



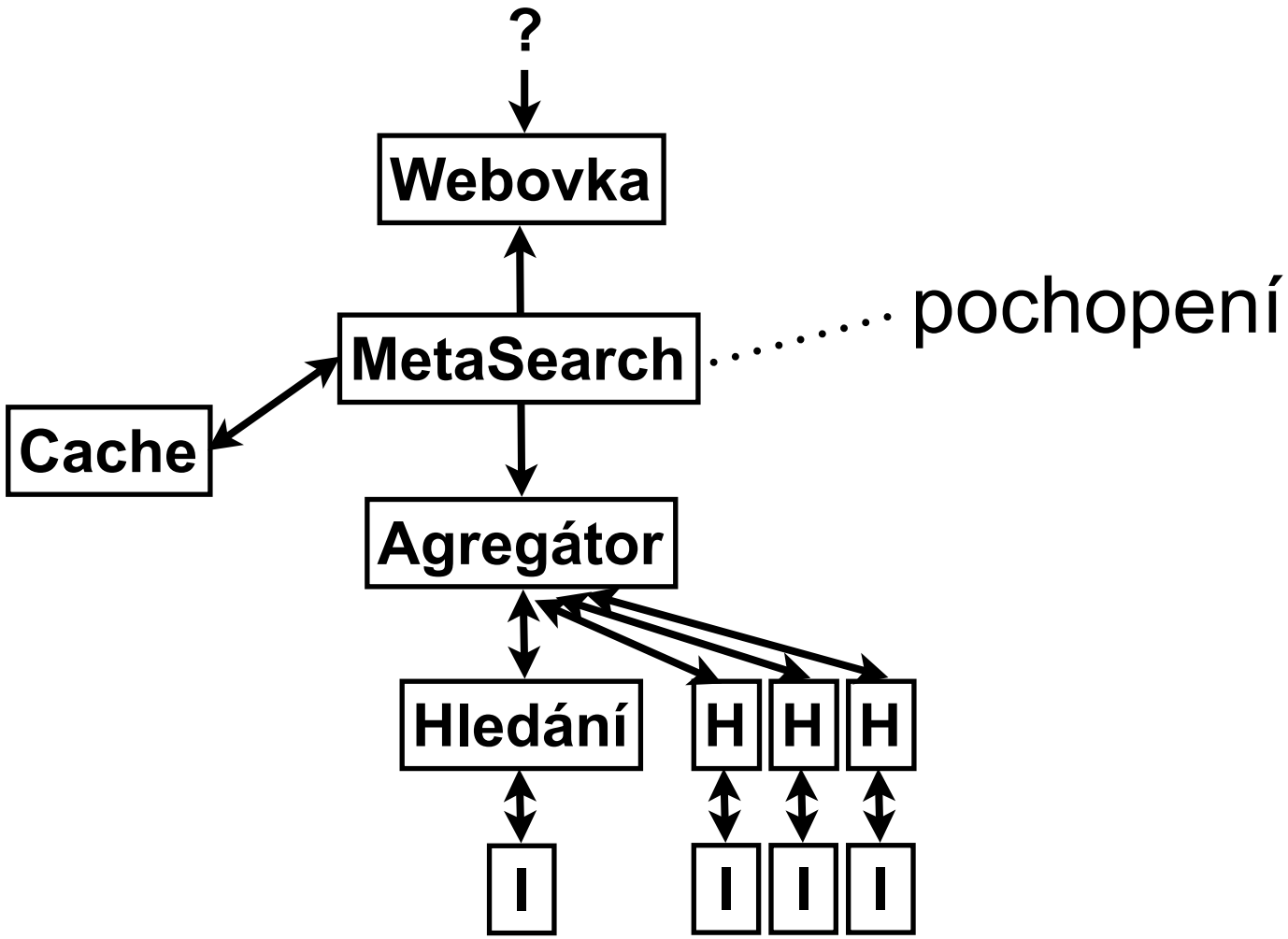
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



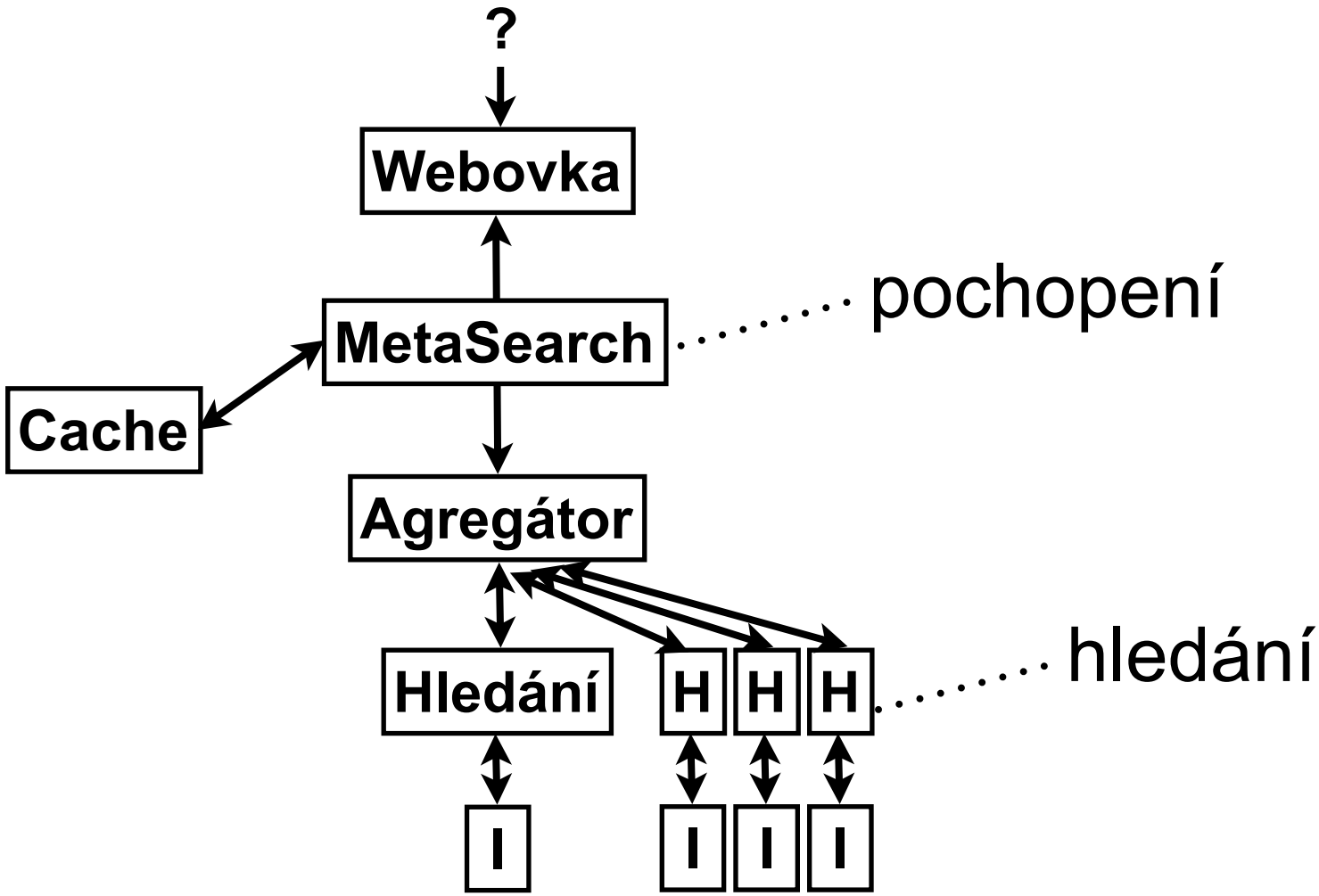
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



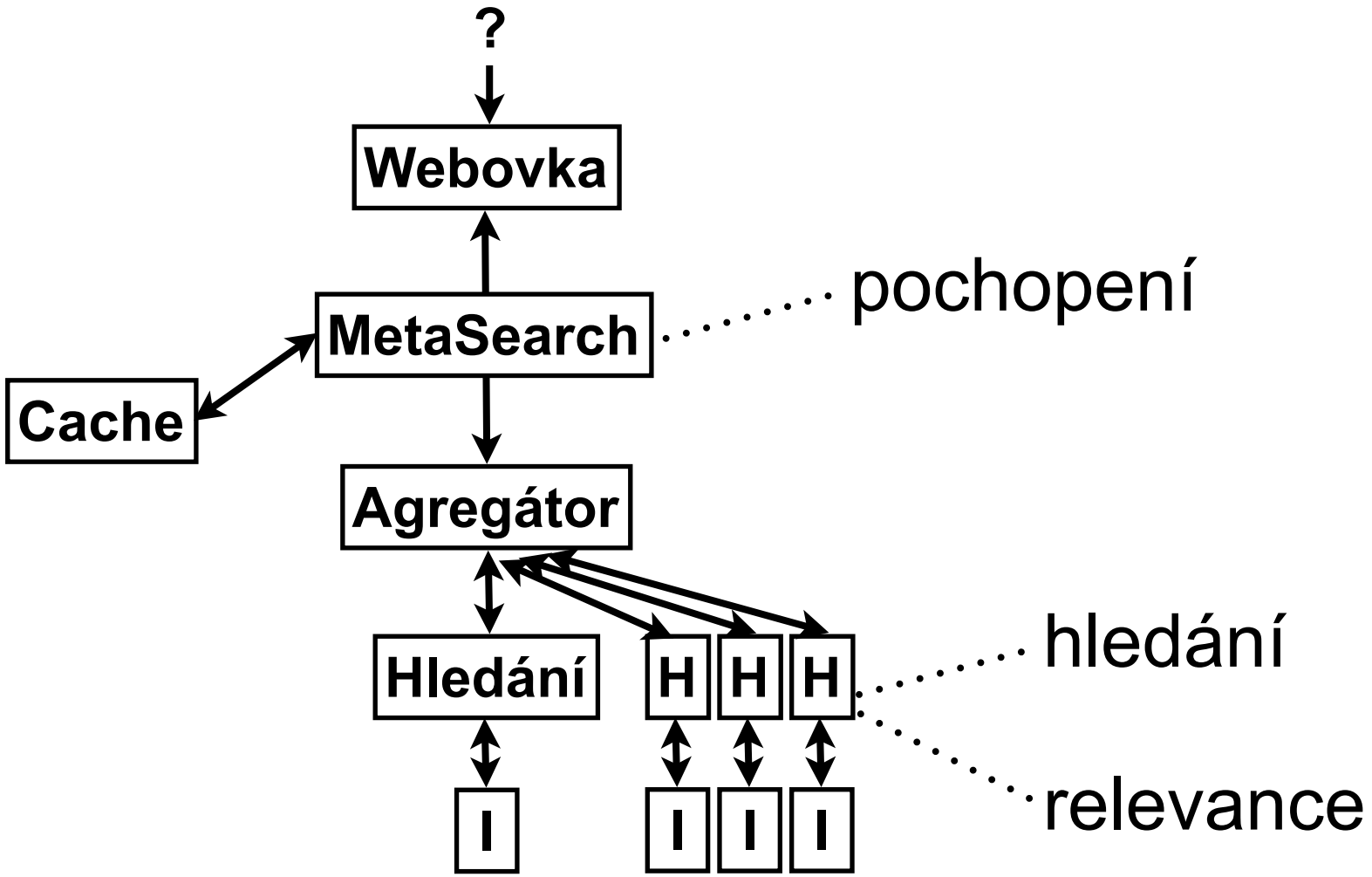
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



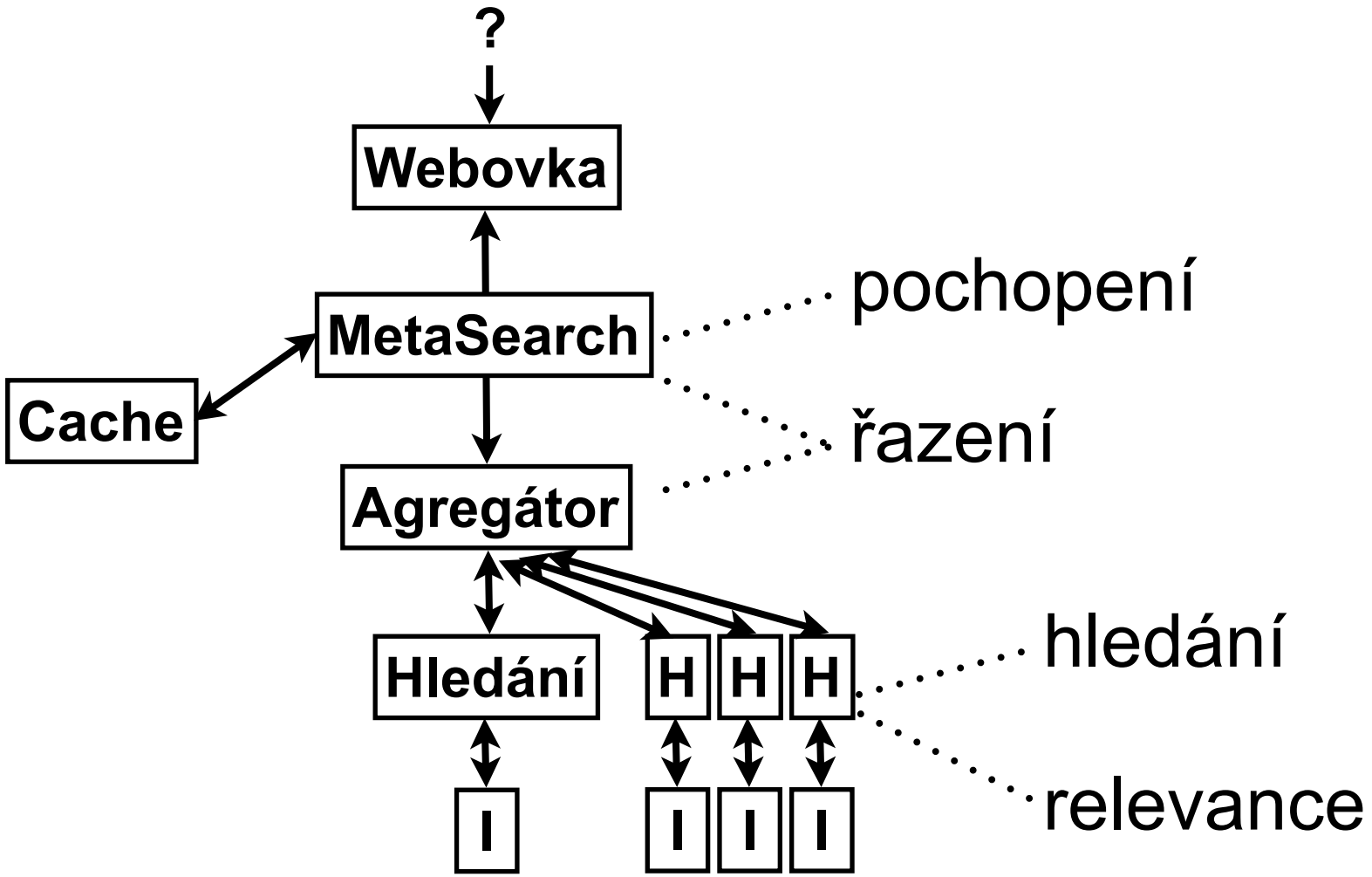
SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



SEZNAM.CZ
...najdu tam, co hledám

Výdej - architektura



SEZNAM.CZ
...najdu tam, co hledám

Přepis dotazu

- Tokenizace
  - “Firefox, Opera?” - [Firefox, Opera]
  - “seznam.cz” - [seznam.cz]
  - “C/C++” - [C/C++]

Přepis dotazu

- Tokenizace
  - “Firefox, Opera?” - [Firefox, Opera]
  - “seznam.cz” - [seznam.cz]
  - “C/C++” - [C/C++]
- Oháčkování

Přepis dotazu

- Tokenizace
  - “Firefox, Opera?” - [Firefox, Opera]
  - “seznam.cz” - [seznam.cz]
  - “C/C++” - [C/C++]
- Oháčkování
- Stop slova
  - “bar **v** Brně” - [bar, Brně]
  - “**jak** vyloupit banku” - [vyloupit, banku]

Přepis dotazu

- Tokenizace
  - “Firefox, Opera?” - [Firefox, Opera]
  - “seznam.cz” - [seznam.cz]
  - “C/C++” - [C/C++]
- Oháčkování
- Stop slova
  - “bar **v** Brně” - [bar, Brně]
  - “**jak** vyloupit banku” - [vyloupit, banku]
- Operátory
  - “+jak”, “recept -brokolice”, “site: .cz”

SEZNAM.CZ
...najdu tam, co hledám

Přepis dotazu

- Kolokace
  - “vlakové nádraží Brno” - [[vlakové, nádraží], Brno]

SEZNAM.CZ
...najdu tam, co hledám

Přepis dotazu

- Kolokace
  - “vlakové nádraží Brno” - [[vlakové, nádraží], Brno]
- Zkratky
  - “ČT” - [ČT **or** [Česká televize]]
  - “Česká televize” - [[Česká televize] **or** ČT]
  - “vše o Praze” vs. “vše v Praze” ?

Přepis dotazu

- Kolokace
  - “vlakové nádraží Brno” - [[vlakové, nádraží], Brno]
- Zkratky
  - “ČT” - [ČT **or** [Česká televize]]
  - “Česká televize” - [[Česká televize] **or** ČT]
  - “vše o Praze” vs. “vše v Praze” ?
- Spojování a rozdělení slov
  - “babybox” - [babybox **or** [baby, box]]
  - “baby box” - [[baby, box] **or** babybox]

Přepis dotazu

- Kolokace
  - “vlakové nádraží Brno” - [[vlakové, nádraží], Brno]
- Zkratky
  - “ČT” - [ČT **or** [Česká televize]]
  - “Česká televize” - [[Česká televize] **or** ČT]
  - “vše o Praze” vs. “vše v Praze” ?
- Spojování a rozdělení slov
  - “babybox” - [babybox **or** [baby, box]]
  - “baby box” - [[baby, box] **or** babybox]
- Čísla
  - “2. světová válka”

SEZNAM.CZ
...najdu tam, co hledám

Strom dotazu

- AND, OR uzly
- váha, proximita
- další atributy

Strom dotazu

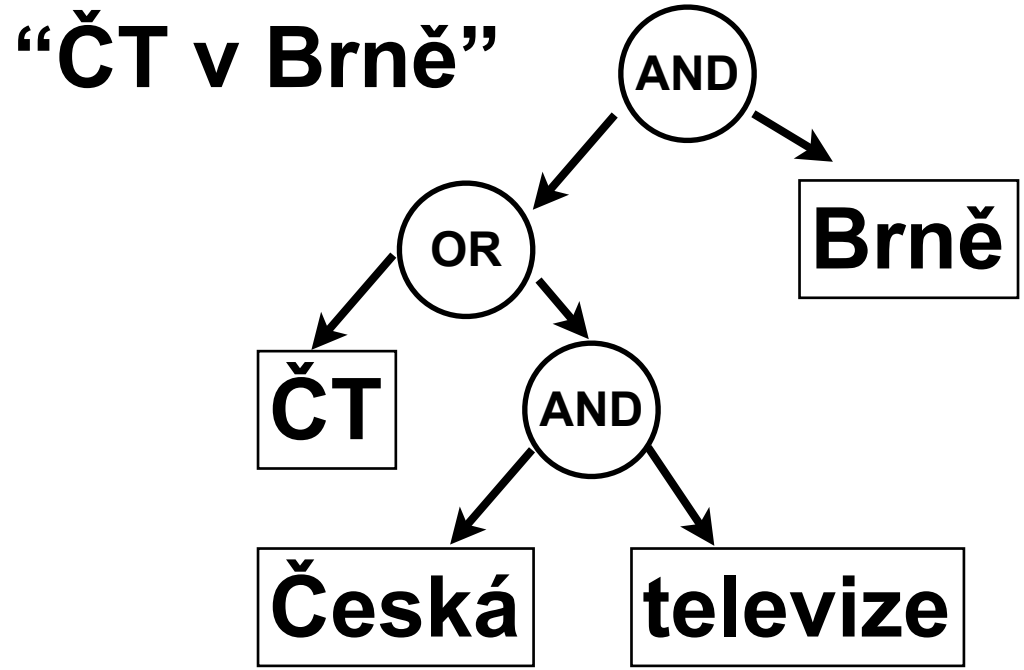
- AND, OR uzly
- váha, proximita
- další atributy

“ČT v Brně”

SEZNAM.CZ
...najdu tam, co hledám

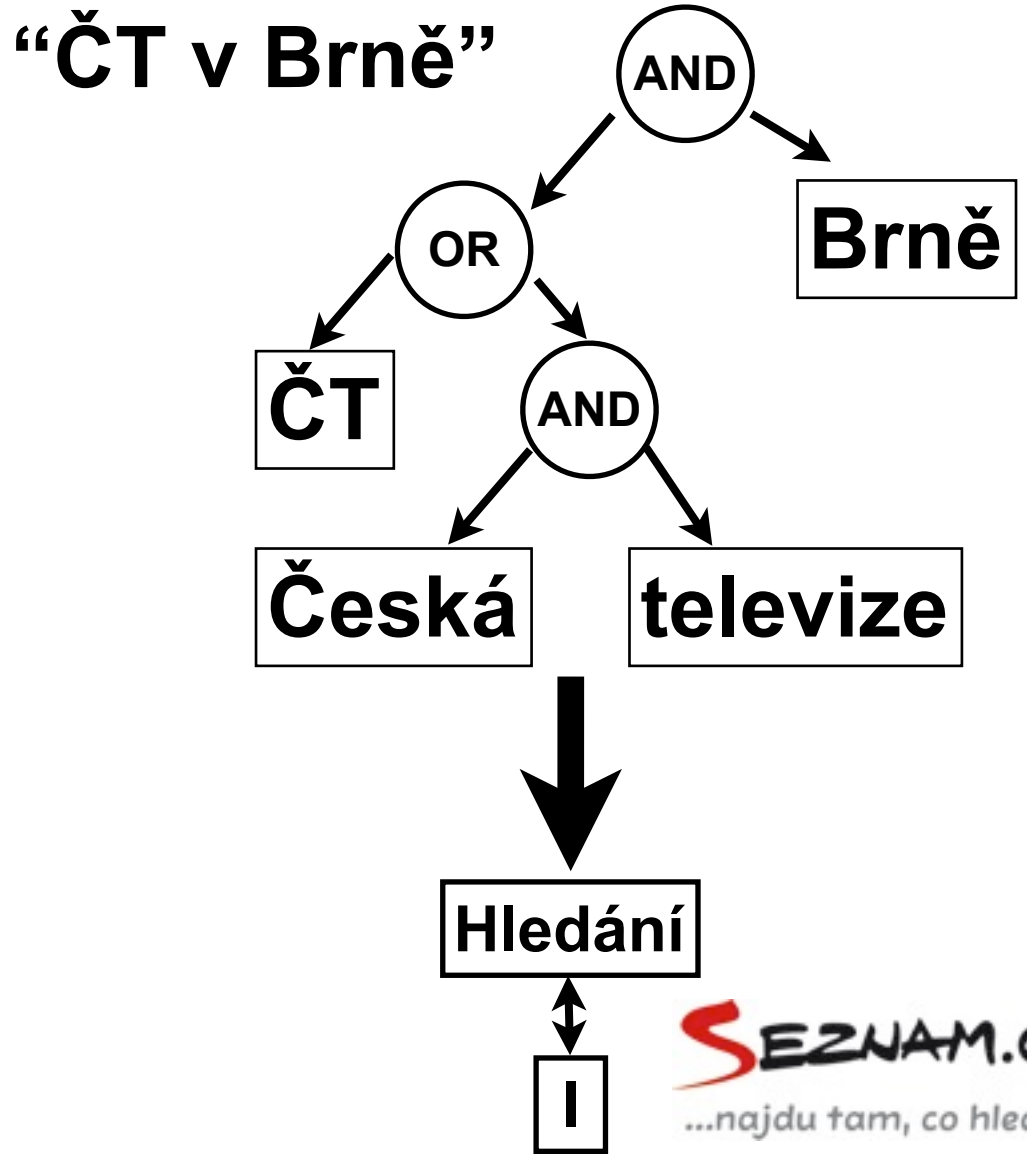
Strom dotazu

- AND, OR uzly
- váha, proximita
- další atributy



Strom dotazu

- AND, OR uzly
- váha, proximita
- další atributy



Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma

Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma

ryba

SEZNAM.CZ
...najdu tam, co hledám

Přepis dotazu - čeština

- základní pojmy
  - term **rybou**
  - lexém
  - lemma **ryba**

SEZNAM.CZ
...najdu tam, co hledám

Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma

rybou

rybám

ryba

SEZNAM.CZ
...najdu tam, co hledám

Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma

rybou

rybám

ryba

ryby

SEZNAM.CZ
...najdu tam, co hledám

Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma

rybou

rybě

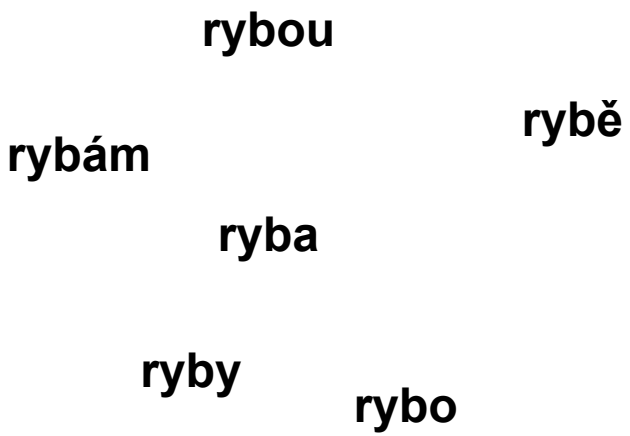
rybám

ryba

ryby

Přepis dotazu - čeština

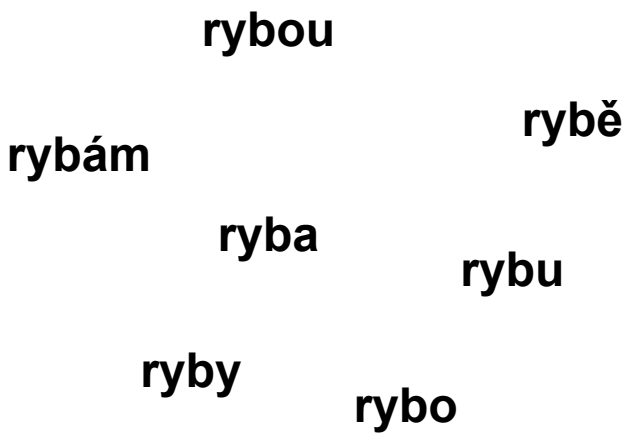
- základní pojmy
  - term
  - lexém
  - lemma



SEZNAM.CZ
...najdu tam, co hledám

Přepis dotazu - čeština

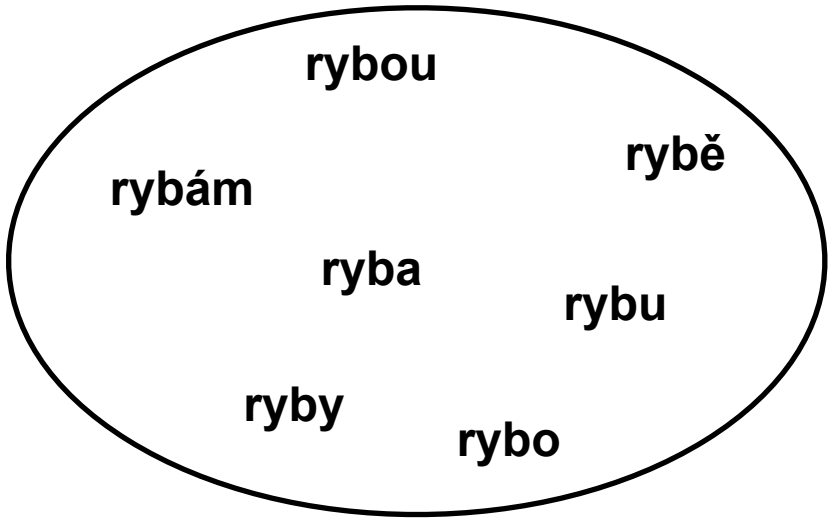
- základní pojmy
  - term
  - lexém
  - lemma



SEZNAM.CZ
...najdu tam, co hledám

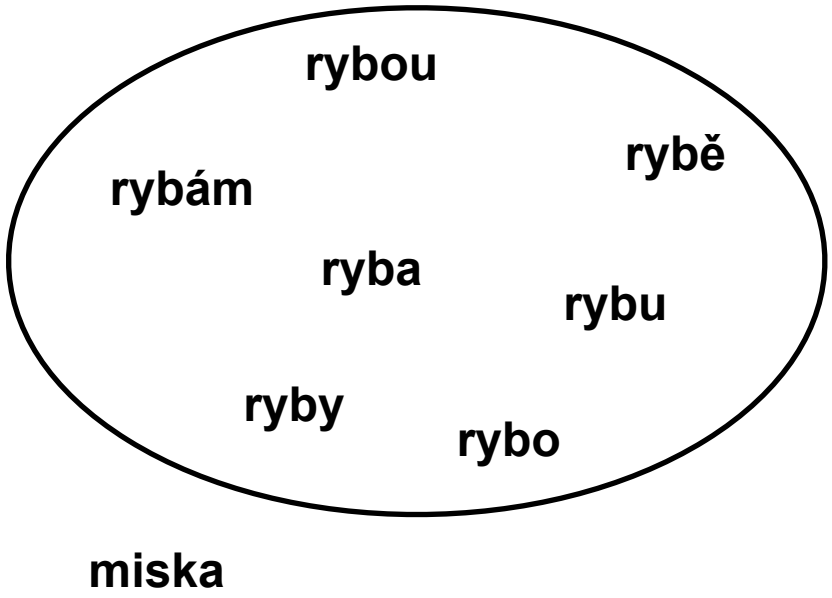
Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma



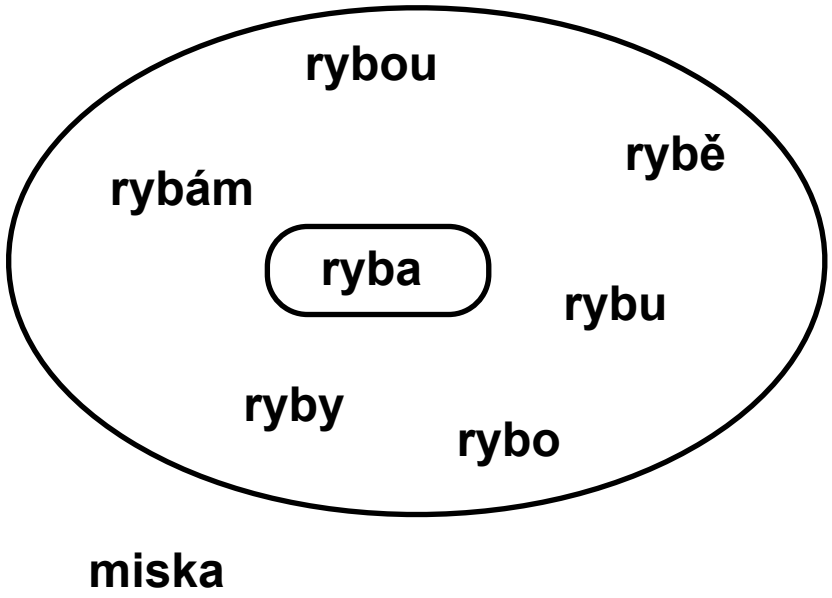
Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma



Přepis dotazu - čeština

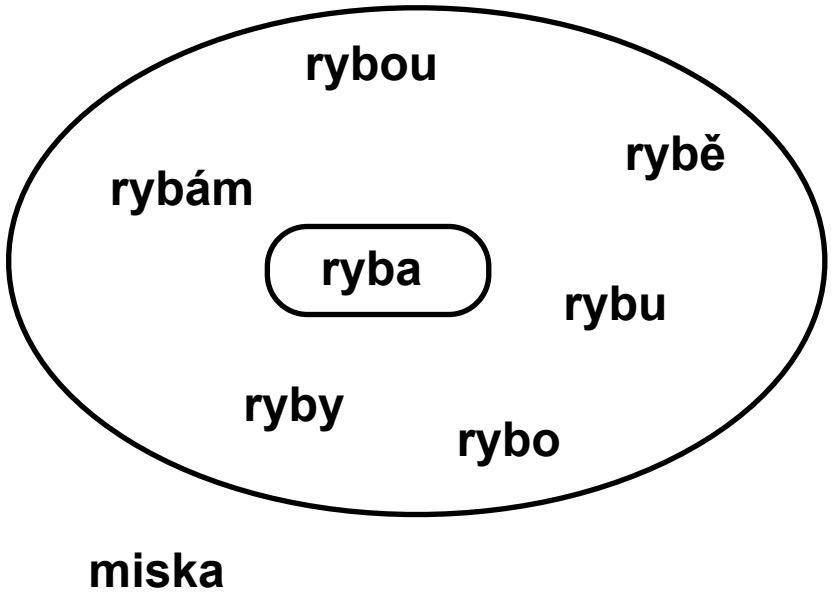
- základní pojmy
  - term
  - lexém
  - lemma



SEZNAM.CZ
...najdu tam, co hledám

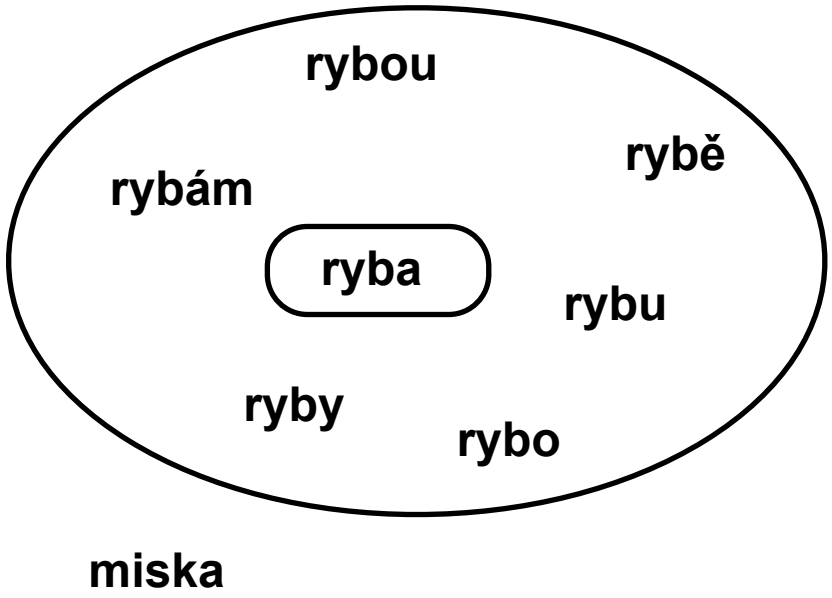
Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma
- Lemmatizace



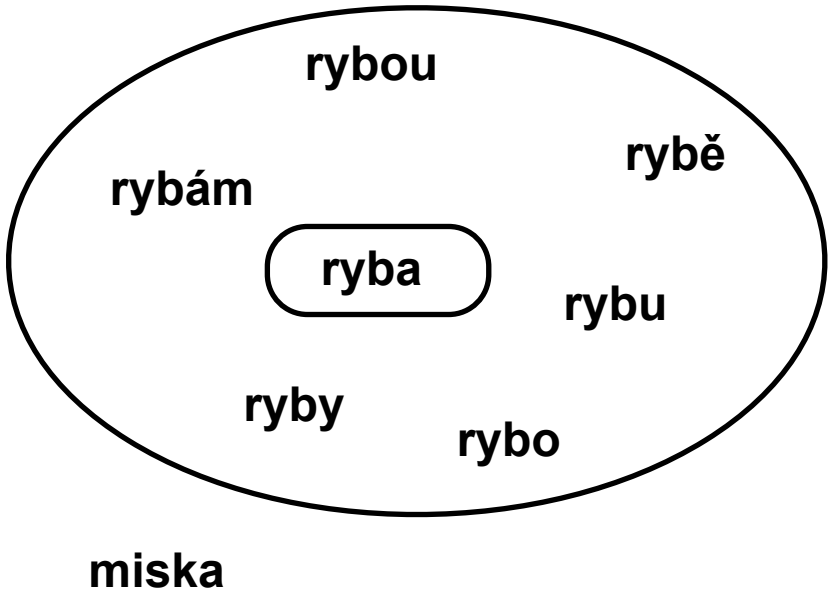
Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma
- Lemmatizace
  - term → lemma



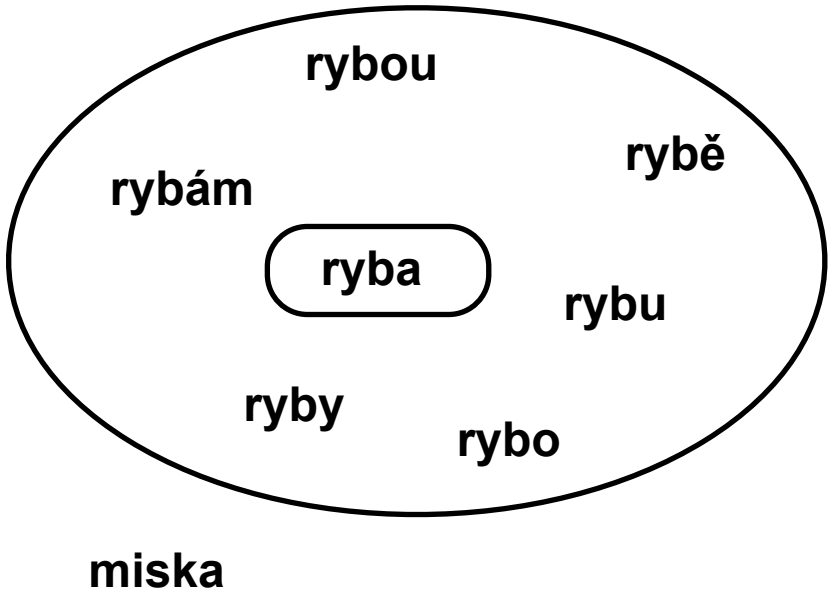
Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma
- Lemmatizace
  - term → lemma
  - lemma → termy



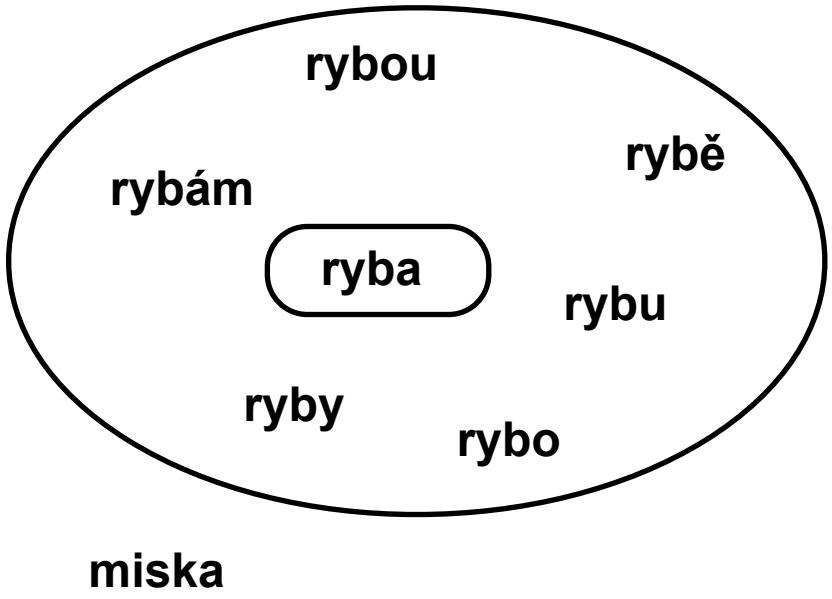
Přepis dotazu - čeština

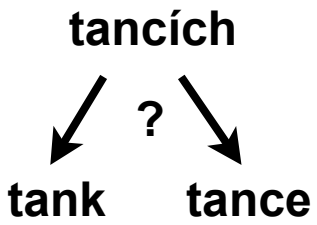
- základní pojmy
  - term
  - lexém
  - lemma
- Lemmatizace
  - term → lemma
  - lemma → termy
  - desambiguace



Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma
- Lemmatizace
  - term → lemma
  - lemma → termy
  - desambiguace

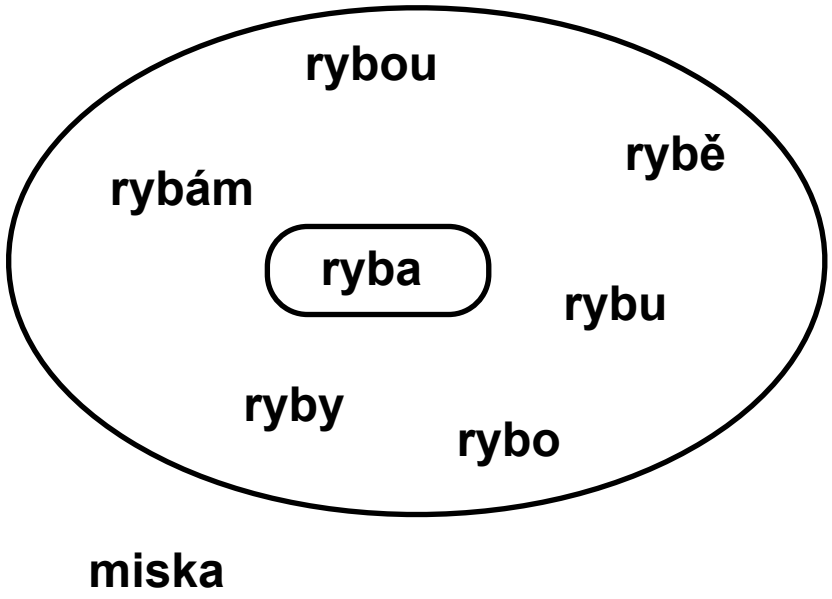


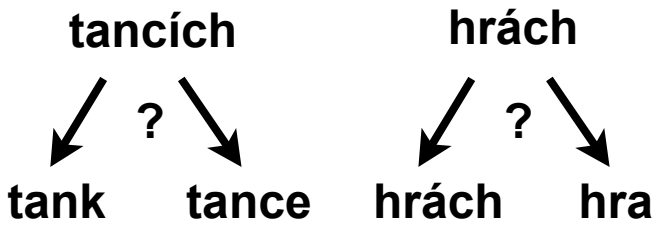


SEZNAM.CZ
...najdu tam, co hledám

Přepis dotazu - čeština

- základní pojmy
  - term
  - lexém
  - lemma
- Lemmatizace
  - term → lemma
  - lemma → termy
  - desambiguace





SEZNAM.CZ
...najdu tam, co hledám

Čeština a indexace

- Indexace lemmat

SEZNAM.CZ
...najdu tam, co hledám

Čeština a indexace

- Indexace lemmat
  - termy dokumentů převést na lemmata

Čeština a indexace

- Indexace lemmat
  - termy dokumentů převést na lemmata
  - lemmata do indexu

Čeština a indexace

- Indexace lemmat
  - termy dokumentů převést na lemmata
  - lemmata do indexu
  - dotaz se převádí na lemmata

Čeština a indexace

- Indexace lemmat
  - termy dokumentů převést na lemmata
  - lemmata do indexu
  - dotaz se převádí na lemmata
  - problémy

Čeština a indexace

- Indexace lemmat
  - termy dokumentů převést na lemmata
  - lemmata do indexu
  - dotaz se převádí na lemmata
  - problémy
    - velké indexy

Čeština a indexace

- Indexace lemmat
  - termy dokumentů převést na lemmata
  - lemmata do indexu
  - dotaz se převádí na lemmata
  - problémy
    - velké indexy
    - nejednoznačnost

Čeština a indexace

- Indexace lemmat
  - termy dokumentů převést na lemmata
  - lemmata do indexu
  - dotaz se převádí na lemmata
  - problémy
    - velké indexy
    - nejednoznačnost
    - různé jazyky

Čeština a indexace

- Indexace lemmat
  - termy dokumentů převést na lemmata
  - lemmata do indexu
  - dotaz se převádí na lemmata
  - problémy
    - velké indexy
    - nejednoznačnost
    - různé jazyky
    - přesné znění

Čeština a indexace

- Indexace termů

SEZNAM.CZ
...najdu tam, co hledám

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu

SEZNAM.CZ
...najdu tam, co hledám

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu
  - “OR” uzel s tvary (+váhy)

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu
  - “OR” uzel s tvary (+váhy)
  - výhody

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu
  - “OR” uzel s tvary (+váhy)
  - výhody
    - menší indexy

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu
  - “OR” uzel s tvary (+váhy)
  - výhody
    - menší indexy
    - částečné řešení desambiguace

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu
  - “OR” uzel s tvary (+váhy)
  - výhody
    - menší indexy
    - částečné řešení desambiguace
    - cizojazyčné expanze

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu
  - “OR” uzel s tvary (+váhy)
  - výhody
    - menší indexy
    - částečné řešení desambiguace
    - cizojazyčné expanze
    - přesné znění

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu
  - “OR” uzel s tvary (+váhy)
  - výhody
    - menší indexy
    - částečné řešení desambiguace
    - cizojazyčné expanze
    - přesné znění
    - operativní možnosti

SEZNAM.CZ
...najdu tam, co hledám

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu
  - “OR” uzel s tvary (+váhy)
  - výhody
    - menší indexy
    - částečné řešení desambiguace
    - cizojazyčné expanze
    - přesné znění
    - operativní možnosti
  - co to neřeší

SEZNAM.CZ
...najdu tam, co hledám

Čeština a indexace

- Indexace termů
  - termy dokumentů přímo do indexu
  - “vyskloňuji” při přepisování dotazu
  - “OR” uzel s tvary (+váhy)
  - výhody
    - menší indexy
    - částečné řešení desambiguace
    - cizojazyčné expanze
    - přesné znění
    - operativní možnosti
  - co to neřeší
    - “kniha o německých tancích”, “ženu holí stroj”

Řazení výsledků

- Vstup: signály

2	1	6	1
1	2	2	1
3	0	5	1
4	1	3	1

Řazení výsledků

- Vstup: signály

Dotaz: “auto”

2	1	6	1
1	2	2	1
3	0	5	1
4	1	3	1

Řazení výsledků

- Vstup: signály

Dotaz: “auto”

doc	v textu	v titulku	PR	slov dotazu	pořadí
A	2	1	6	1	
B	1	2	2	1	
C	3	0	5	1	
D	4	1	3	1	

Řazení výsledků

- Vstup: signály

Dotaz: “auto”

doc	v textu	v titulku	PR	slov dotazu	pořadí
A	2				
B	1				
C	3				
D	4				

Řazení výsledků

- Vstup: signály

Dotaz: “auto”

doc	v textu	v titulku	PR	slov dotazu	pořadí
A	2	1			
B	1	2			
C	3	0			
D	4	1			

Řazení výsledků

- Vstup: signály

Dotaz: “auto”

doc	v textu	v titulku	PR	slov dotazu	pořadí
A	2	1	6		
B	1	2	2		
C	3	0	5		
D	4	1	3		

Řazení výsledků

- Vstup: signály

Dotaz: “auto”

doc	v textu	v titulku	PR	slov dotazu	pořadí
A	2	1	6	1	
B	1	2	2	1	
C	3	0	5	1	
D	4	1	3	1	

Řazení výsledků

- Vstup: signály

Dotaz: “auto”

doc	v textu	v titulku	PR	slov dotazu	pořadí
A	2	1	6	1	
B	1	2	2	1	
C	3	0	5	1	
D	4	1	3	1	

- Lineární kombinace

Řazení výsledků

- Vstup: signály

Dotaz: “auto”

doc	v textu	v titulku	PR	slov dotazu	pořadí
A	2	1	6	1	
B	1	2	2	1	
C	3	0	5	1	
D	4	1	3	1	

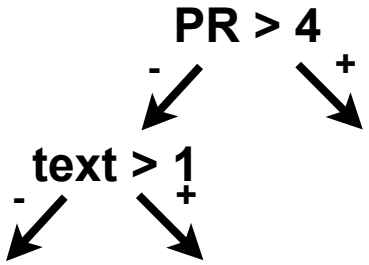
- Lineární kombinace
- Regresní rozhodovací stromy

SEZNAM.CZ
...najdu tam, co hledám

Rozhodovací stromy

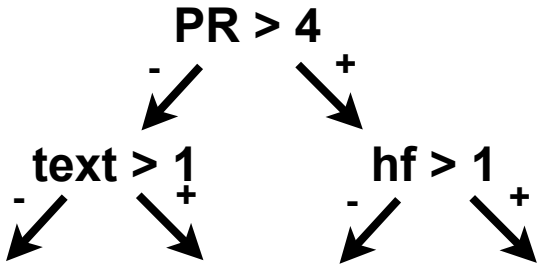
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



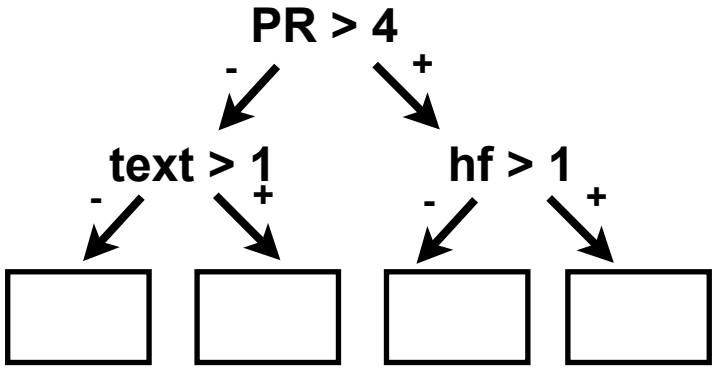
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



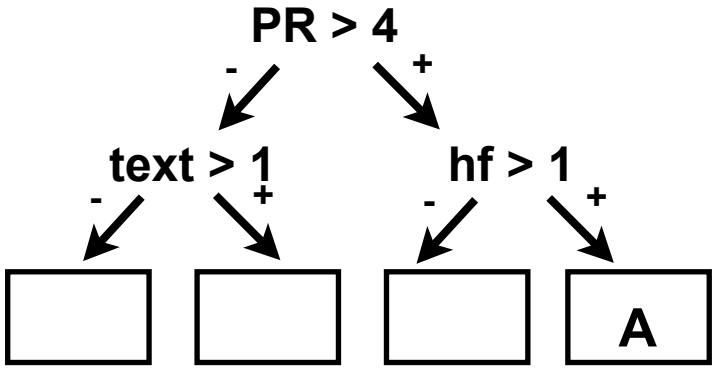
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



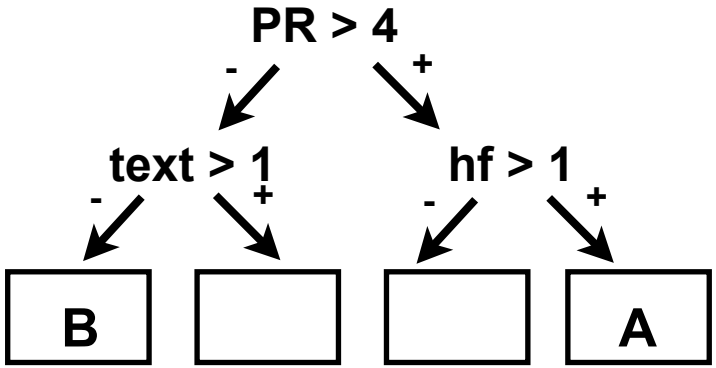
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



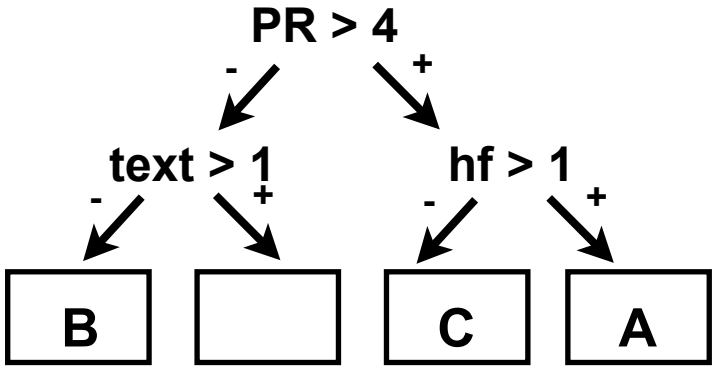
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



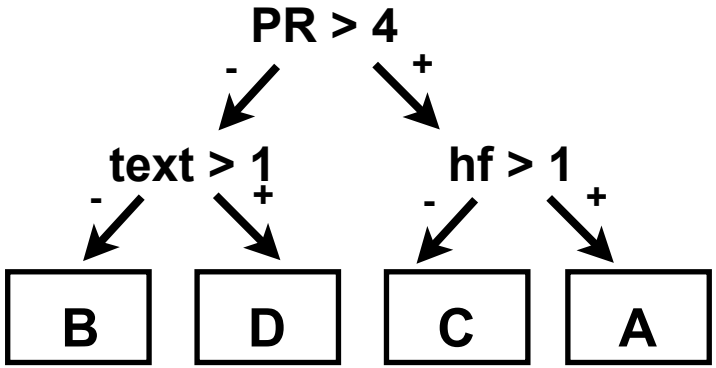
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



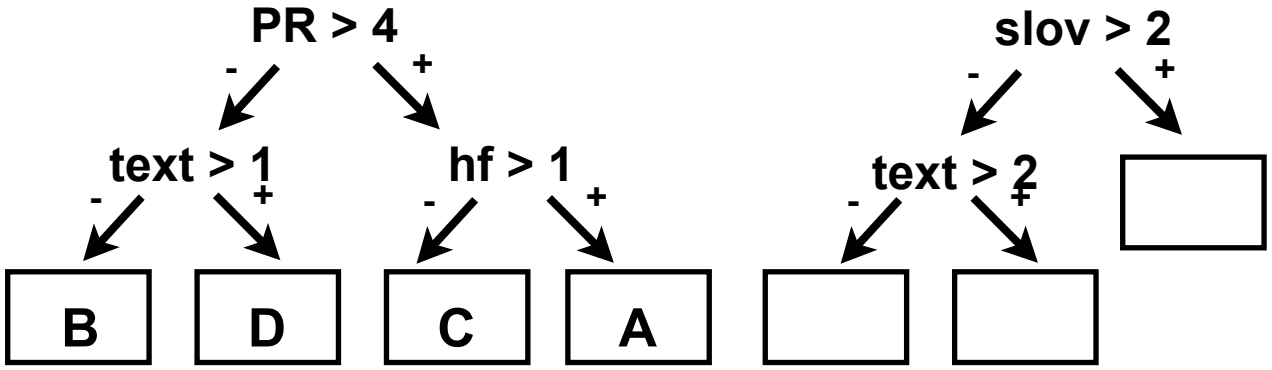
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



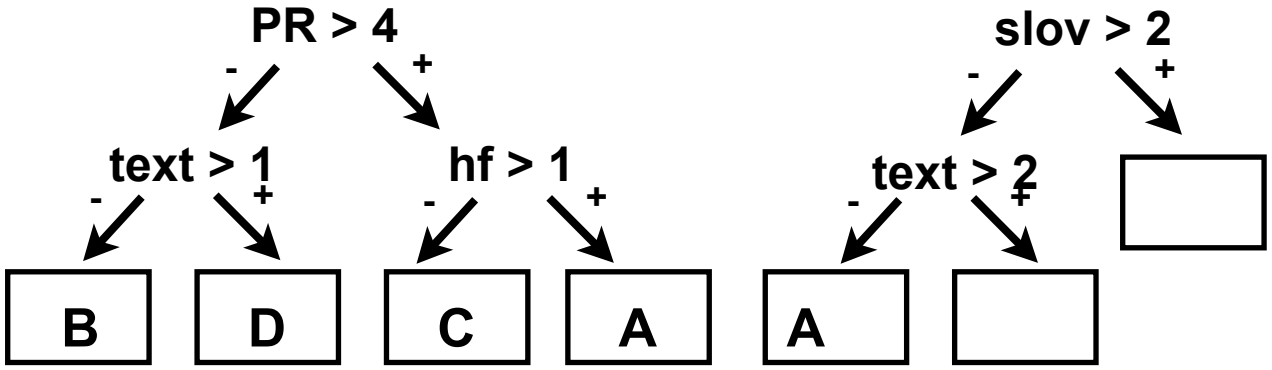
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



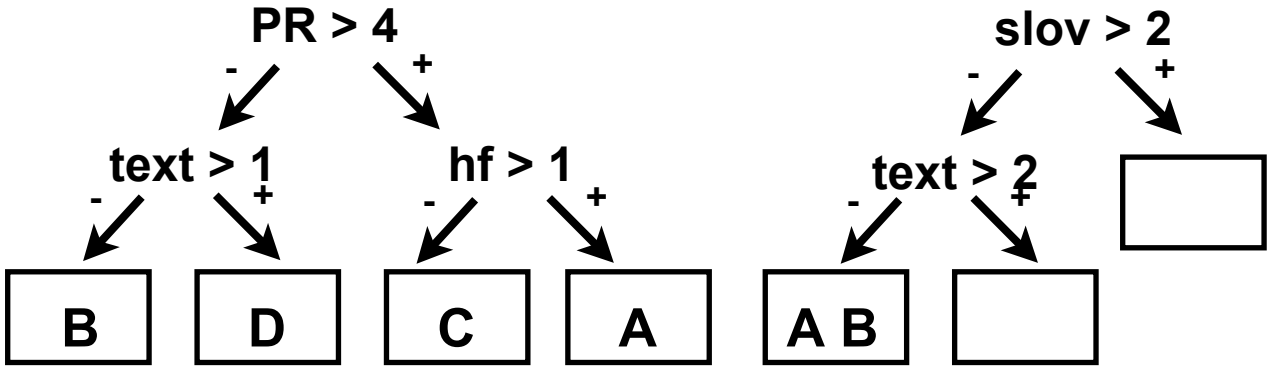
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



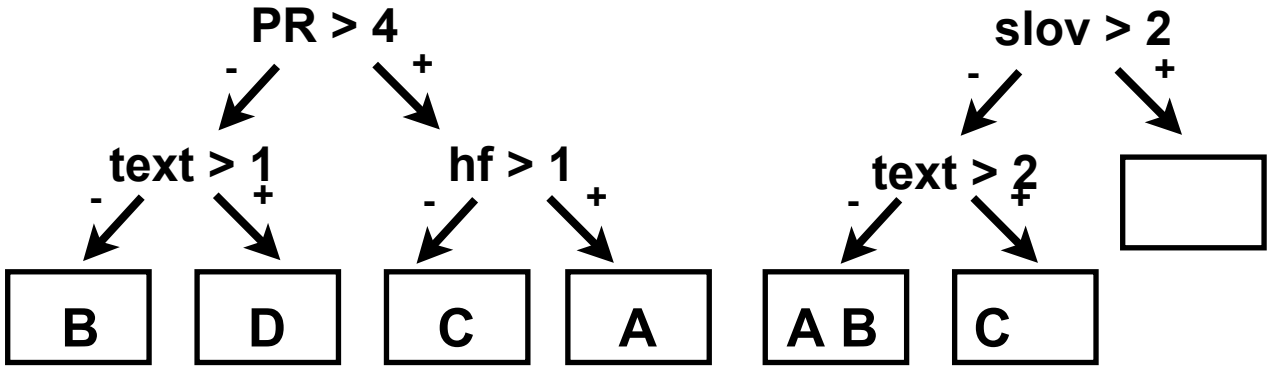
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



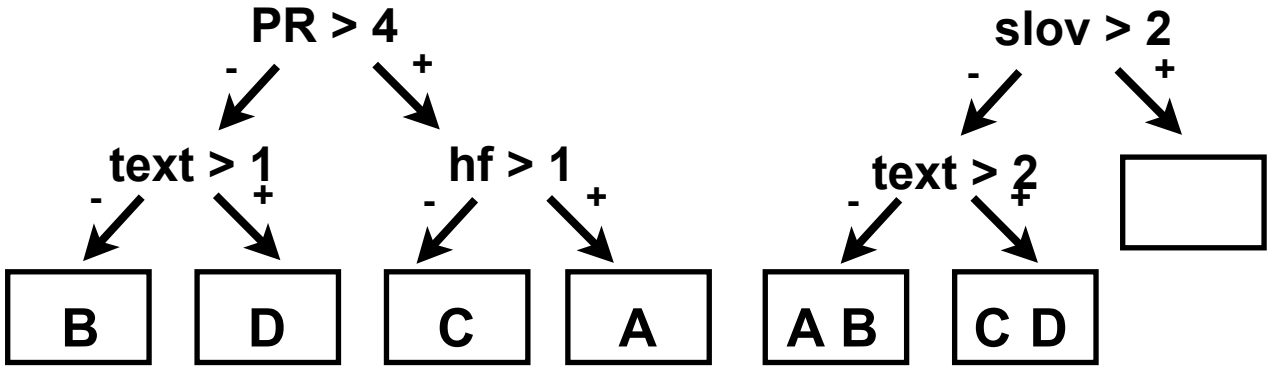
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



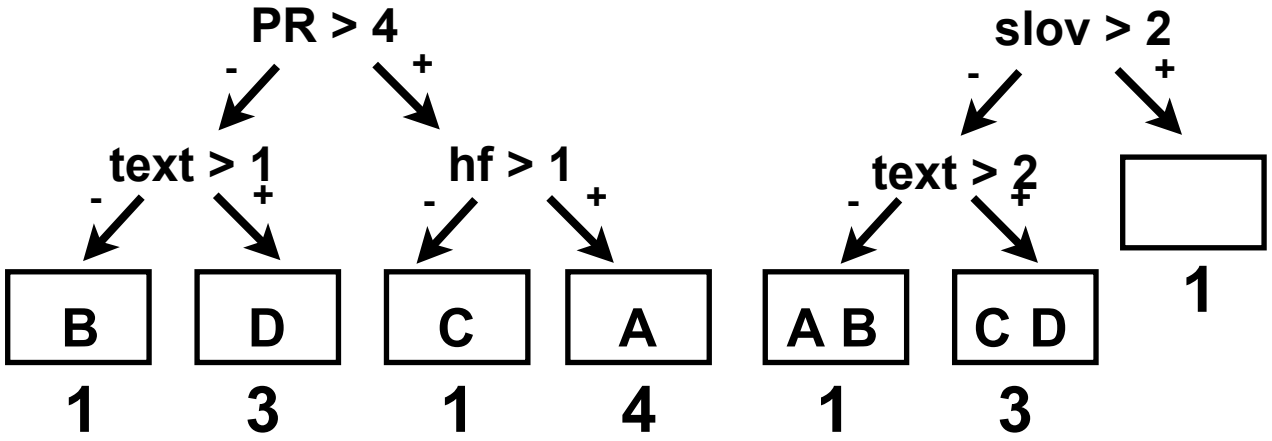
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



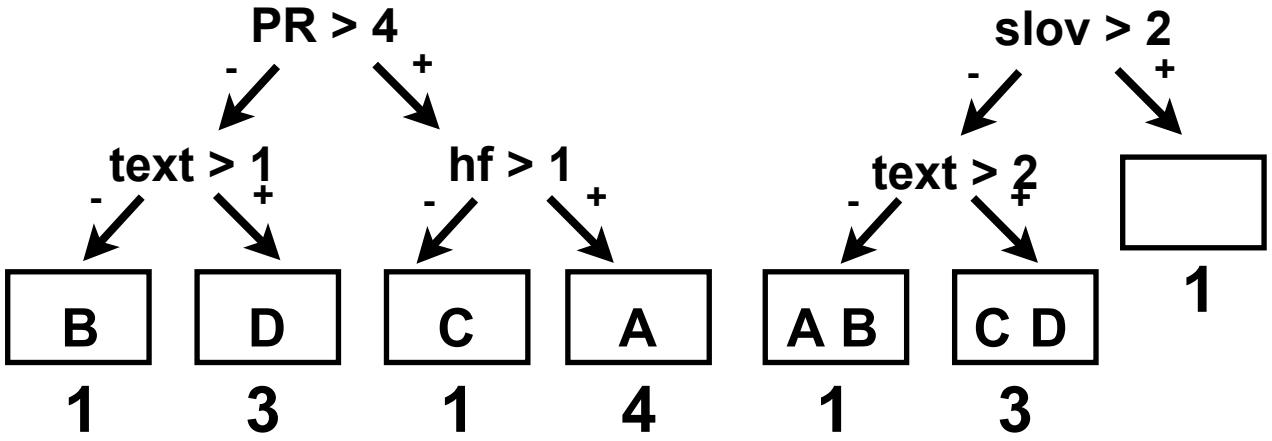
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



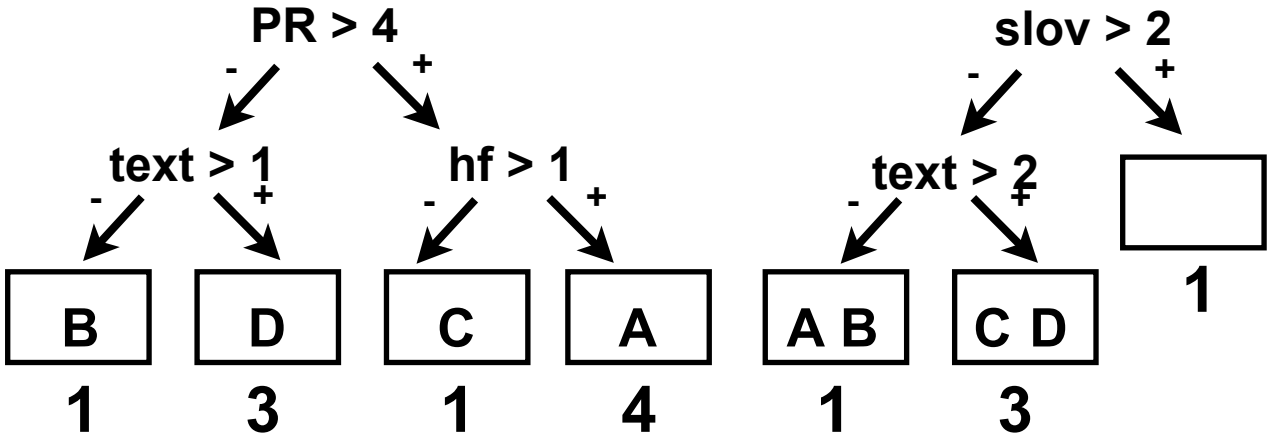
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy



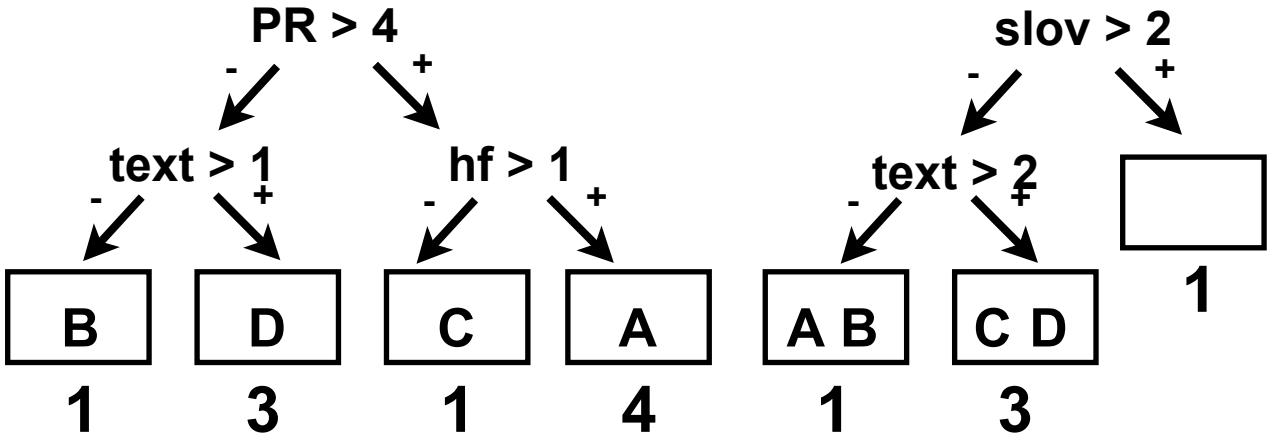
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	
C	3	0	5	1	
D	4	1	3	1	

Rozhodovací stromy



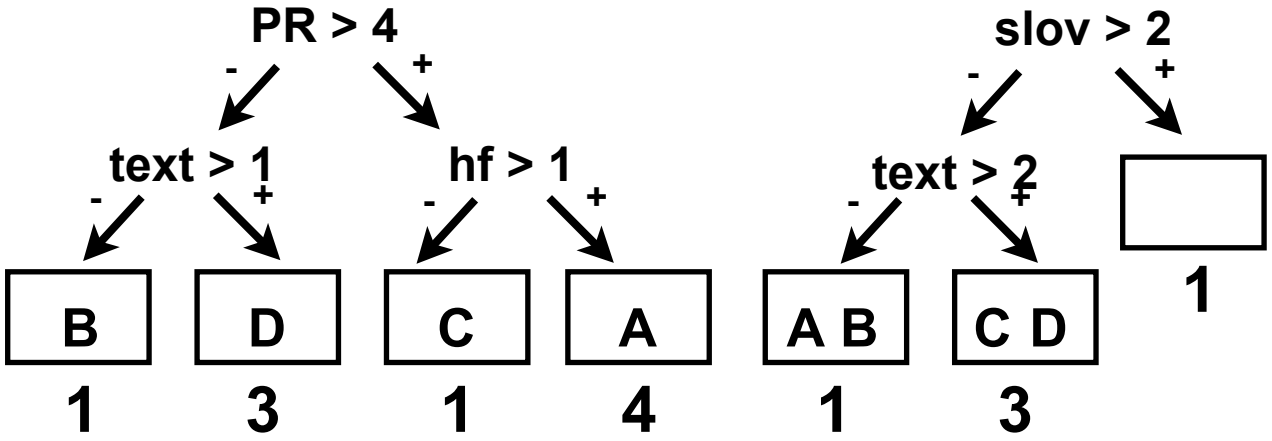
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	
D	4	1	3	1	

Rozhodovací stromy



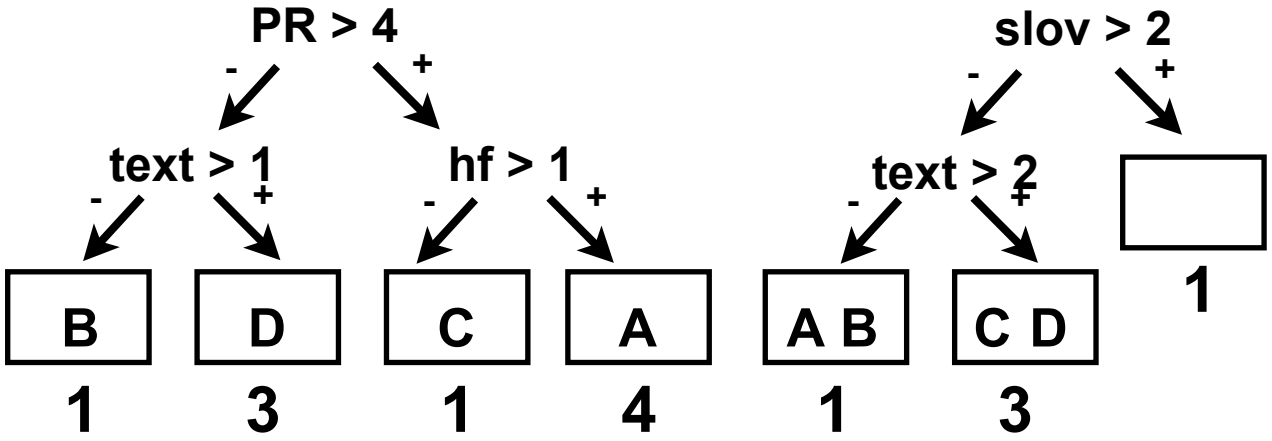
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	

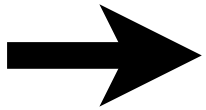
Rozhodovací stromy



doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Rozhodovací stromy

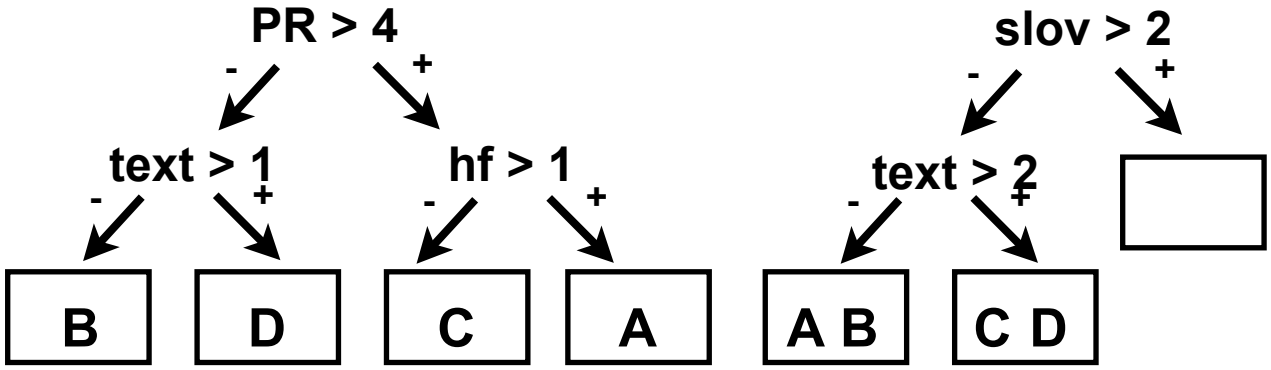




1. D
2. A
3. C
4. B

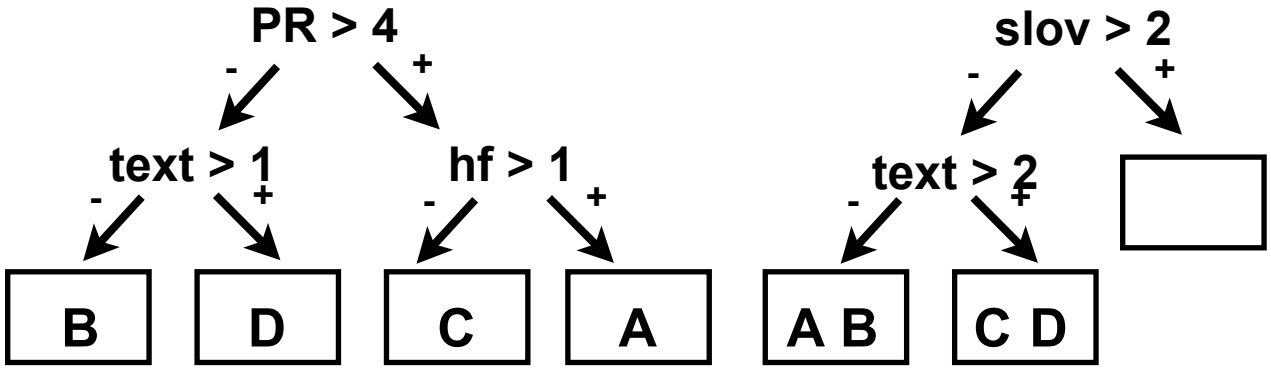
doc	text	hf	PR	slov	pořadí
A	2	1	6	1	5
B	1	2	2	1	2
C	3	0	5	1	4
D	4	1	3	1	6

Stavění stromů



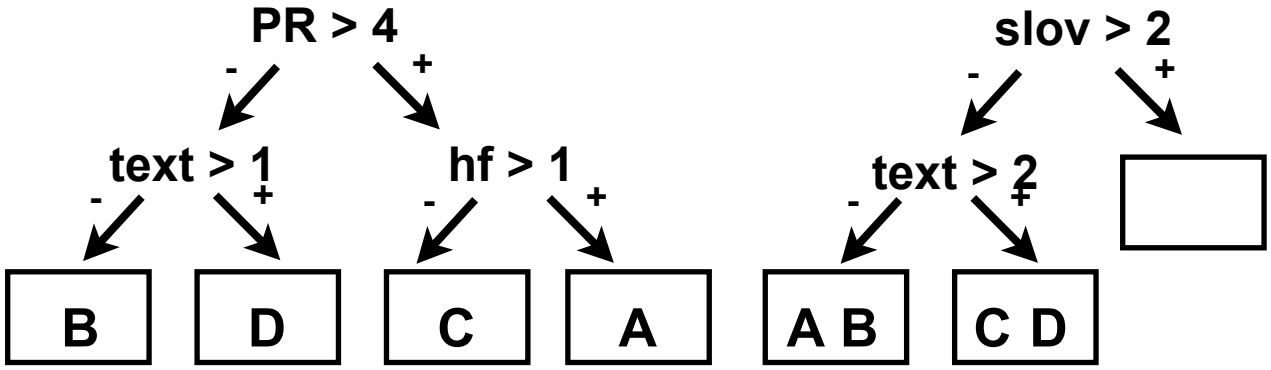
doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1		6.5
B	1	2	2	1		3.5
C	3	0	5	1		5.5
D	4	1	3	1		8.5

Stavění stromů



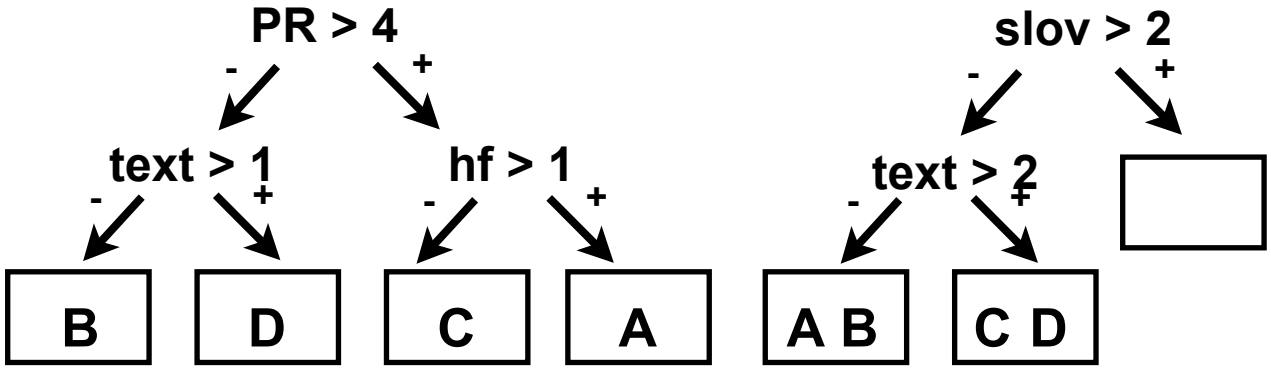
doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1		3.5
C	3	0	5	1		5.5
D	4	1	3	1		8.5

Stavění stromů



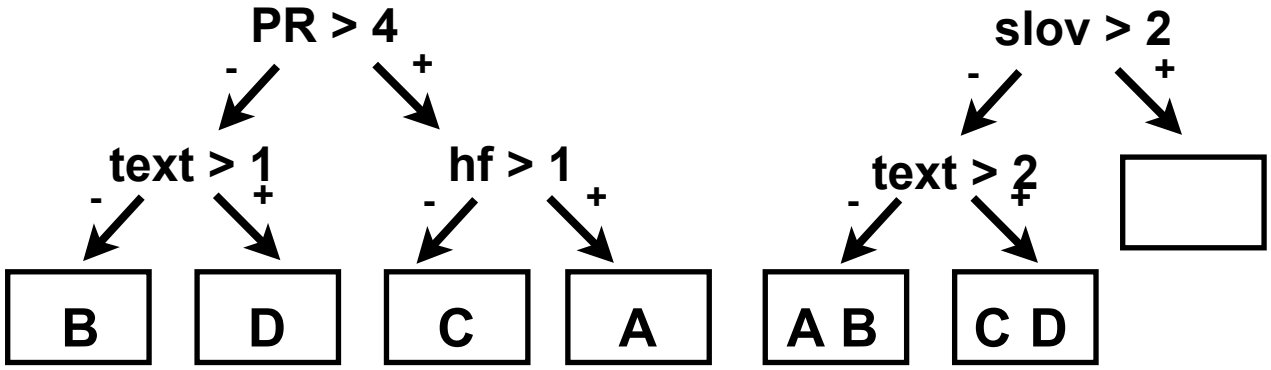
doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1	1	3.5
C	3	0	5	1		5.5
D	4	1	3	1		8.5

Stavění stromů



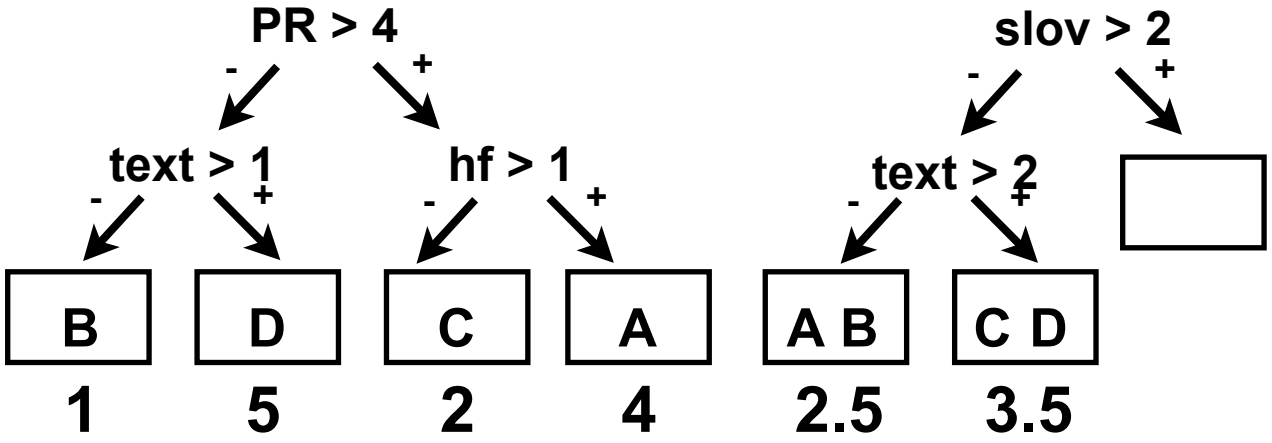
doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1	1	3.5
C	3	0	5	1	2	5.5
D	4	1	3	1		8.5

Stavění stromů



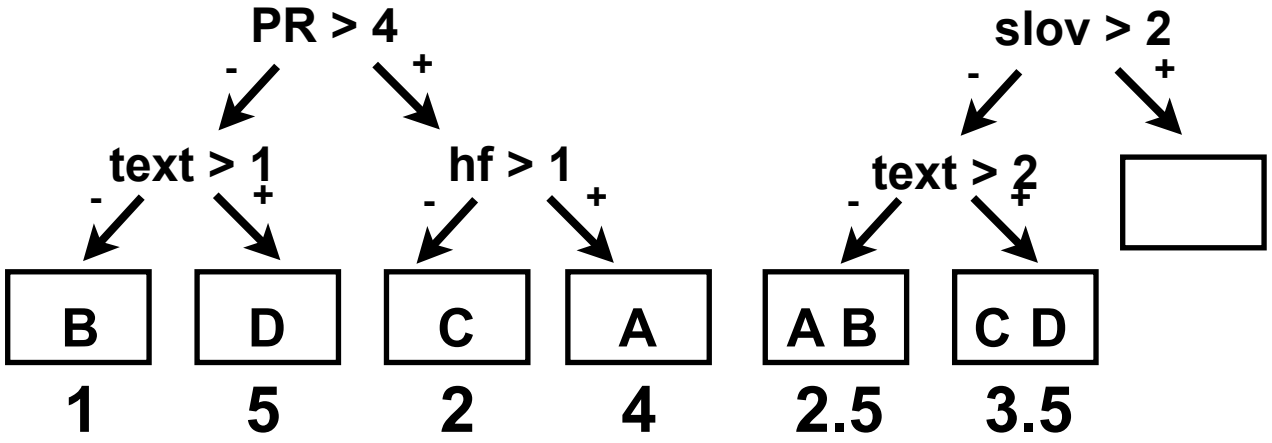
doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1	1	3.5
C	3	0	5	1	2	5.5
D	4	1	3	1	5	8.5

Stavění stromů



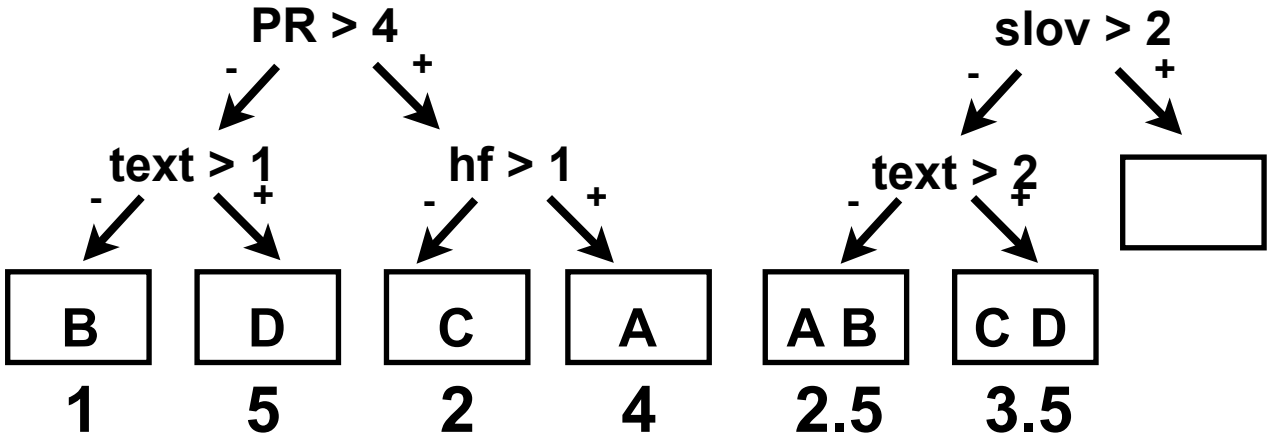
doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1	1	3.5
C	3	0	5	1	2	5.5
D	4	1	3	1	5	8.5

Stavění stromů



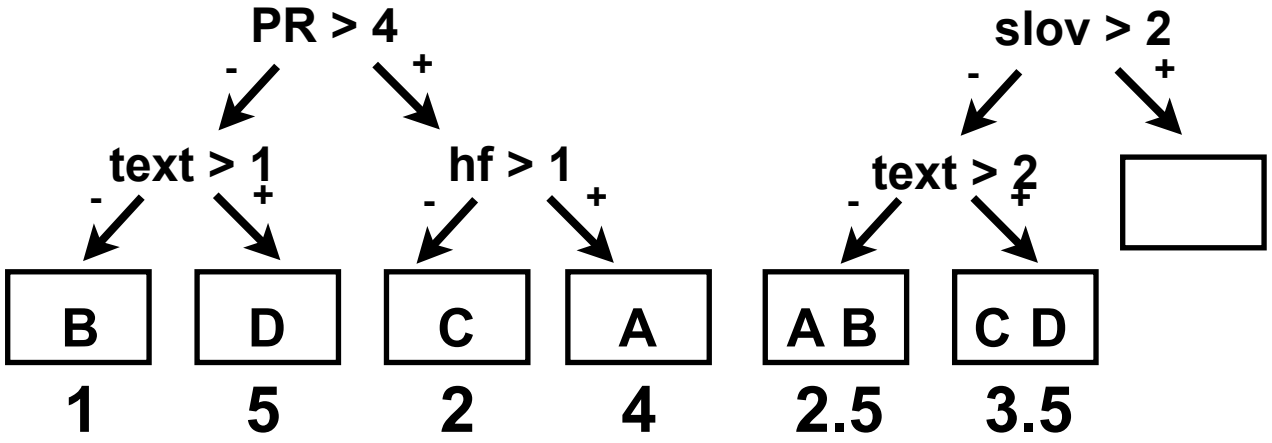
doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1	1	
C	3	0	5	1	2	
D	4	1	3	1	5	

Stavění stromů



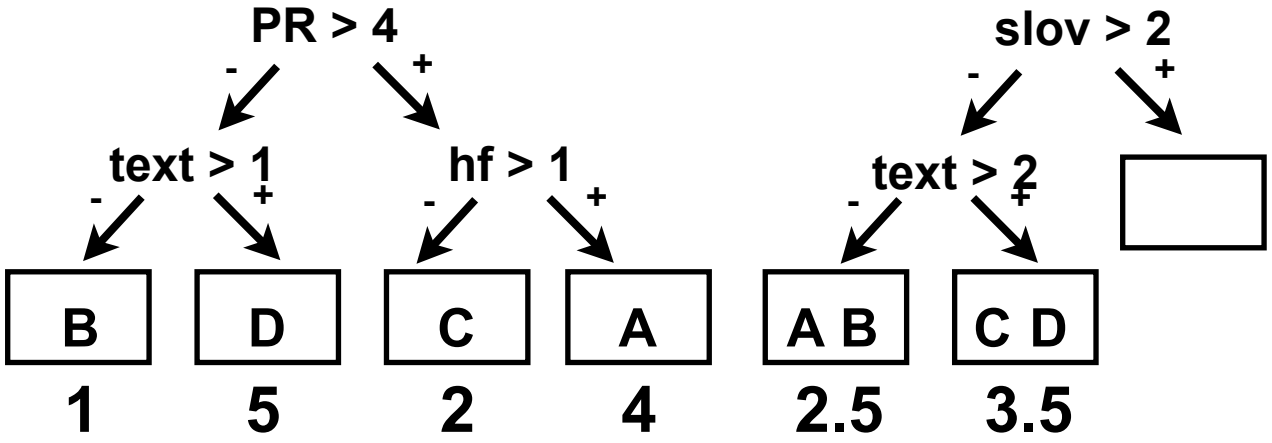
doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1	1	3.5
C	3	0	5	1	2	
D	4	1	3	1	5	

Stavění stromů



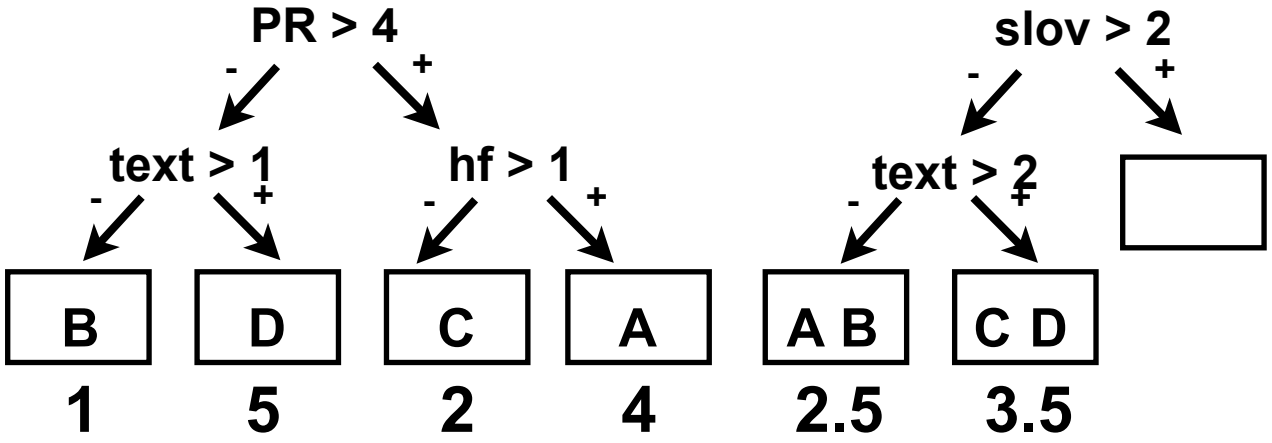
doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1	1	3.5
C	3	0	5	1	2	5.5
D	4	1	3	1	5	

Stavění stromů



doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1	1	3.5
C	3	0	5	1	2	5.5
D	4	1	3	1	5	8.5

Stavění stromů



doc	text	hf	PR	slov	ruční	pořadí
A	2	1	6	1	4	6.5
B	1	2	2	1	1	3.5
C	3	0	5	1	2	5.5
D	4	1	3	1	5	8.5
E	2	0	7	1		4.5

Dodatky

- Překlepy
- Související dotazy
- Zpětná vazba
- Oháčkování 4x jinak
- Rozhodovací stromy
  - spam, porno

SEZNAM.CZ
...najdu tam, co hledám

Seznam.cz Vyhledávání

- Jak to začalo
- Co máme
- Co používáme
- Co řesíme

SEZNAM.CZ

Jak to začalo

- Rok = 2005
- Stroje = 11
- Lidé = 4

SEZNAM.CZ

Co máme

- Celkem 70 lidí
- 32 programátorů
- 16 produkt manažerů
- 12 výzkumníků
- 6 administrátorů

- 100 brigádníků

Co máme

- Celkem 500 strojů
- 300 vyhledávání
- 150 robot
- 50 vývoj a výzkum
- 2 serverovny

SEZNAM.CZ

Co máme

- 500TB dat
- 50M dokumentů denně
- Rychlostí několik GBit/s
- Hledáme v 800M dokumentech
- 350 až 500 dotazů za sekundu

SEZNAM.CZ

Co používáme

APACHE HBASE

python

debian

Java

OpenVZ
Linux Containers

C++

hadoop

SEZNAM.CZ

Co řesíme

- “kniha o německých tancích” ?
- Relevance
- Relevance
- Relevance
- Relevance
- …

SEZNAM.CZ

Děkuji za pozornost…

SEZNAM.CZ

...najdu tam, co hledám


SEZNAM.CZ
...najdu tam, co hledám